

**Robust Knowledge Acquisition
in Answering Information-seeking Questions
At Scale**

by

Ehsan Kamaloo

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science
University of Alberta

© Ehsan Kamaloo, 2022

Abstract

Answering information-seeking question involves retrieving relevant documents from a massive haystack of unstructured text corpora. This dissertation aims at building question answering (QA) systems that can be deployed in the wild where incoming questions may be noisy and their distribution inevitably shifts from that of the training data. At a high level, we attempt to tackle three distinct problems arising in real-world scenarios: from the modelling perspective, how to build robust and scalable QA models, and how to acquire knowledge that is useful for fulfilling questions from text, and from the evaluation perspective, how to reliably evaluate retrieval-based QA models.

Towards this goal, we first study the problem of adapting classical IR models for QA tasks. For this purpose, we investigate one of the basic and salient linguistic features in text, the relationship between the ordering of words in an answer passage and that of a question. In particular, we present a sparse retrieval model that treats n -grams as single compound terms to represent local word order. Second, our focus shifts to the generalizability of QA models via data augmentation. To this end, we design a sample-efficient data augmentation framework, inspired by adversarial training methods, that makes QA models robust to distribution shift. Third, we present a novel knowledge acquisition method that can be helpful in addressing ambiguity in questions. In particular, we aim at automatically deriving meta-information about the spatial grounding of location mentions in text. Our method does not require any supervision and leverages the structural interactions between the mentions in a document. Finally, we focus on the reliability of evaluation benchmarks in information-

seeking QA. Specifically, we highlight that existing benchmarks are heavily skewed toward passage-level information. Our analysis paves the way for designing future benchmarks that can better reflect the true performance of QA models.

Overall, in pursuit of achieving genuine human-level QA systems that can be readily used in real-world applications, the present thesis highlights the key requirements of knowledge acquisition, robustness to distribution shift, scalability, and reliable evaluation.

Preface

The central chapters of this thesis are based on papers that are either published or that is currently under review. In particular, Chapter 3 and Chapter 4 are written based on papers that are published in conference proceedings [1, 2, 3]. Chapter 5 is based on a paper that is currently under review [4]. Finally, Chapter 2 is an original contribution to this thesis.

- [1] E. Kamaloo, M. Rezagholizadeh, P. Passban, and A. Ghodsi, “Not Far Away, Not So Close: Sample Efficient Nearest Neighbour Data Augmentation via MiniMax,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Dublin, Ireland: Association for Computational Linguistics, Aug. 2021, pp. 3522–3533, DOI: 10.18653/v1/2021.findings-acl.309. [Online]. Available: <https://aclanthology.org/2021.findings-acl.309>.

- [2] E. Kamaloo, M. Rezagholizadeh, and A. Ghodsi, “When Chosen Wisely, More Data Is What You Need: A Universal Sample-Efficient Strategy For Data Augmentation,” in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, Jul. 2022, pp. 1048–1062, DOI: 10.18653/v1/2022.findings-acl.84. [Online]. Available: <https://aclanthology.org/2022.findings-acl.84>.

- [3] E. Kamaloo and D. Rafiei, “A Coherent Unsupervised model for Toponym Resolution,” in *Proceedings of the 2018 World Wide Web Conference*, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 1287–1296, DOI: 10.1145/3178876.3186027.

- [4] E. Kamaloo, C. L. Clarke, and D. Rafiei, “Document-level Reasoning: A Hidden Challenge in Open-Domain Question Answering Benchmarks,” Under review.

To my extraordinary mom for her unconditional support

Acknowledgements

When I embarked on this journey, I had no clue about what lies ahead for me. I experienced the ebb and flow of research that helped me grow personally and professionally. However, without the help of many people, I would not have been able to make it to the end. I feel indebted to those who have helped me along the way.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Davood Rafiei, for his unwavering support, and his invaluable mentorship. He gave me the opportunity to pursue my passion for research even though I had been away from academia for a few years. He has patiently endured my mistakes and has never thought twice to dedicate his time to me whenever I needed it. I cannot imagine doing this thesis without Davood's help and support.

I wish to thank my PhD committee members: Dr. Osmar Zaiane, Dr. Greg Kondrak, and Dr. Di Niu for taking the time to attend several checkpoints and providing valuable comments. I also thank my external examiner, Dr. Jian-Yun Nie, for thoughtful questions and insightful feedback.

I am incredibly grateful to Dr. Mehdi Rezaghlizadeh and Dr. Charles Clarke whose mentorship, knowledge, and contributions were instrumental in advancing my research including and beyond this thesis. I cannot overstate how much I owe to their insights that were critical to my development as a researcher.

My sincere gratitude goes to Nouha Dziri, my amazing friend and my wonderful collaborator. Nouha and I started working together on a course project, but our collaboration has never ended ever since. Nouha no doubt had a profound influence on me to be a better researcher. Her attention to details and her writing skills were a

great source of inspiration. Her boundless energy and enthusiasm kept me motivated and helped me push my limits.

I was fortunate to be surrounded by friends who made this journey joyful: Amirhossein Nazari, Yadollah Yaghoobzadeh, Saeed Sarabchi, Sepehr Kazemian, Athar Mahmoodinejad, Shadan Golestan-Irani, Anahita Doosti, Mohammad Motallebi, Behdad Bakhshinategh, Kamyar Khodamoradi, Sanket Kumar Singh, Megha Panda, Mahdi Rahmani, and Amir Samani. I have fond memories of Pouneh Gorji and Arash Pourzarabi who tragically were among the victims of Flight PS752. I also thank folks at the University of Alberta Swim Club that was my sanctuary during these years and in particular, Myron Liew, Ramin Mousavi, Devon Christoffel, Won-Yong Song, Brent Bester, and Calvin Howard for the unforgettable memories.

Lastly, I am eternally indebted to my Mom, Parvin, my brother, Amir, and my sister, Azadeh, who have always been there for me no matter what. My family encouraged me to pursue my passion and taught me perseverance. They have been my lifelong pillars of support.

Table of Contents

1	Introduction	1
1.1	Thesis Statement	2
1.1.1	Answering Information-seeking Questions	3
1.1.2	Knowledge Acquisition	4
1.1.3	Robustness	5
1.2	Key Contributions	5
1.3	Dissertation Layout	8
2	Revisiting Local Word Order in Retrieval	9
2.1	Introduction	9
2.2	Related Work	11
2.3	Local Word Order-Aware Retrieval	12
2.4	Experiments	14
2.4.1	Experimental Setup	14
2.4.2	Passage Ranking	15
2.4.3	End-to-End Results	16
2.4.4	Ablation Study	17
2.4.5	Document Ranking	19
2.5	Conclusions	20
3	Robustness Via Sample-Efficient Data Augmentation	22
3.1	Introduction	22

3.2	Related Work	24
3.2.1	Task-agnostic DA in NLP	24
3.2.2	Task-aware DA in NLP	25
3.2.3	DA for KD	25
3.3	Methodology	26
3.3.1	General DA for Single Networks	28
3.3.2	DA for Teacher-Student (KD)	29
3.4	Experiments	30
3.4.1	Setup	30
3.4.2	GLUE	31
3.4.3	HellaSwag	35
3.4.4	SQuAD	36
3.5	Ablation Study and Discussion	37
3.5.1	Runtime Analysis	38
3.5.2	Effect of Pre-processing Augmented Data	39
3.6	Reproducibility	40
3.6.1	Fine-tuning Details	40
3.6.2	Distillation Details	41
3.7	Conclusion	42
4	An Unsupervised Model for Grounding Location Mentions	44
4.1	Introduction	45
4.2	Problem Definition	47
4.3	The Unsupervised Model	48
4.3.1	Context-Bound Hypotheses	49
4.3.2	Spatial-Hierarchy Sets	54
4.3.3	Context-Hierarchy Fusion	57
4.4	Experiments	57

4.4.1	Datasets	59
4.4.2	Evaluation Metrics	61
4.4.3	Analysis of Context-Bound Hypotheses	62
4.4.4	Fusion Threshold Study	63
4.4.5	Resolution Accuracy	64
4.4.6	Unseen Data Analysis	66
4.5	Related Works	68
4.6	Conclusions	71
5	Document-level Reasoning: A Hidden Challenge in Open-domain QA Benchmarks	72
5.1	Introduction	72
5.2	Related Work	74
5.3	Document-level Reasoning QA Challenge	75
5.3.1	Setup	75
5.3.2	Document retrieval vs. Passage retrieval	76
5.3.3	Data Collection	78
5.4	Experiments	80
5.5	Experimental Setup	81
5.5.1	Predicting Granularity Level of Retrieval	82
5.5.2	End-to-End Results	83
5.5.3	Varying Passage Length	84
5.5.4	Enrich Passages with Additional Context	84
5.6	Conclusion	85
6	Conclusion and Future Work	87
6.1	Summary of Contributions	87
6.2	Future Work	90
6.2.1	Considerations for using models in the wild	90

6.2.2	Scalability and Complex Reasoning	91
6.2.3	Sample Efficiency for Dense Retrieval	91
6.2.4	Data-centric Analysis of QA Datasets	91

Bibliography		93
---------------------	--	-----------

List of Tables

2.1	Statistics of QA benchmarks. $ Q $ denotes the average number of words in questions.	15
2.2	Hits ratio at top- k passages for various retrievers.	16
2.3	Exact-Match (EM) accuracy of various open-domain QA pipelines.	17
2.4	Hits ratio and MRR at top-100 passages for all variants of our sparse retriever. [†] and [‡] indicate statistical significance (p -value < 0.05) over u^- (row 2), unigram retrieval that is a widely-adopted retriever in open-domain QA, and over $u^- + b^-$ (row 5), a DrQA-inspired retriever, respectively.	19
2.5	Hits ratio and MRR at top-10 documents for various retrieval models. BM25 retrievers were run with different sets of unigrams and bigrams, explained in Table 2.4. [†] indicates statistical significance (<i>i.e.</i> McNemar’s test for hits@ k and Student’s t-test for MRR@ k with p -value < 0.05) over u^-	20
3.1	Test results of the distilled experiment on GLUE. The augmentation size is $8x$. Bold and <u>underlined</u> numbers indicate the best and the second best results across the DA methods.	32
3.2	Test result of the standalone experiments on GLUE using RoBERTa _{base}	33
3.3	Dev results of the standalone experiment on GLUE using RoBERTa _{base} . For MMEL, the results are obtained from our implementation.	34

3.4	OOD results for the distilled mode. Bold numbers indicate the best result across DistilRoB models.	35
3.5	OOD results for <i>test</i> settings in the standalone mode. Bold numbers indicate the best result.	35
3.6	OOD results for <i>dev</i> settings in the standalone mode. For MMEL, the results are obtained from our implementation. Bold numbers indicate the best result.	36
3.7	Dev results of the distilled experiment on two downstream tasks.	36
3.8	OOD results for models trained on SQuAD and tested on QA datasets from four different domains [130].	37
3.9	Dev results of self-KD exhibiting the effectiveness of different pre-processing techniques to filter augmented examples on 4 GLUE tasks. β and LP depict a minimum confidence threshold, and label preserving, respectively.	40
3.10	Dev results of self-KD for studying the effect of augmentation size and the selection algorithm for 4 GLUE tasks.	41
3.11	Hyperparameters of DistilRoBERTa on two downstream tasks.	42
3.12	Hyperparameters of DistilRoBERTa on the GLUE benchmark. We used the same configuration for RoBERTa _{base} albeit with a few exceptions marked by (*).	42
4.1	Corpora used in our experiments	60
4.2	Detailed analysis of Context-Bound Hypotheses (CBH) on <i>TR-News</i> dataset	63
4.3	Performance results in <i>GeoTag</i> and <i>Resol</i> experiments on LGL and CLUST. The best results in each category are bolded.	67
4.4	Performance results in <i>GeoTag</i> and <i>Resol</i> experiments on TR-News. The best results in each category are bolded.	67

5.1	Number of questions for which document retrieval surpasses passage retrieval	78
5.2	The breakdown of passage retrieval failures.	81
5.3	Dataset Statistic, constructed from NQ-OPEN, for predicting the granularity level of retrieval.	83
5.4	Exact-match accuracy of our retrievers, paired with FiD [81], on NQ-OPEN.	84

List of Figures

2.1	MRR@ k on NQ-open for 5 variants of our proposed models using BM25. $u_{\rightsquigarrow}^- + b_{\rightsquigarrow}^{*+}$ leads other models.	18
3.1	Illustration of Glitter (from left to right): first, generating augmented samples from different DA techniques; second, forming a pool of samples $X'(i)$; third, evaluating the augmented samples using the --- loss; fourth, filtering the top- k_1 samples based on their corresponding --- ; fifth, updating the parameters of the model by minimizing the task loss $\ell_{\text{task}}(:\theta)$	26
3.2	Runtime Analysis of DA when training $\text{RoBERTa}_{\text{base}}$ using self-KD. The red point signifies Glitter $8x/2x$, except for RTE that is $8x/1x$	38
4.1	Comparative analysis of the test datasets based on location type	61
4.2	F_1 -measure vs. threshold τ for Context-Hierarchy Fusion method on <i>TR-News</i> dataset. At $\tau = 0.55$, CHF achieves the best F_1 -measure on all three corpora.	64
4.3	F_1 -measure of CustomAdaptive trained on CLUST and CHF when overlap ratio varies. CHF yields a better performance than CustomAdaptive when overlap between training data and test data is lower than 60%.	68

5.1	An example question, taken from the Natural Questions-OPEN test set [101], that highlights the importance of document-level reasoning in retrieving passages.	73
5.2	Hits ratio vs. text volume (left), and for the subset that document retrieval performs better (right). Although both passage retrievers outperform document retrieval by a high margin on the full dataset, document retrieval significantly outperforms both passage retrievers on the selected questions.	77
5.3	Manual inspection of 325 questions where document retrieval is superior to passage retrieval over all three open-domain QA benchmarks. .	80
5.4	Hits ratio at volume 10K for various passage lengths on NQ-OPEN and our benchmark.	85

Abbreviations

DA Data Augmentation.

IR Information Retrieval.

KB Knowledge Base.

KD Knowledge Distillation.

LM Language Model.

NLP Natural Language Processing.

NLU Natural Language Understanding.

PLM Pre-trained Language Model.

QA Question Answering.

Chapter 1

Introduction

What makes humans special? “We are ‘rational animals’ pursuing knowledge for its own sake. We live by art and reasoning.” Aristotle answers. We, humans, have an intrinsic desire to know more. Our epistemic curiosity motivates us to ask questions about the world. It is not surprising that shortly after the invention of general-purpose digital computers, researchers sought to build machines that are capable of automatically answering our questions. In the early days of Question Answering (QA), models hinged on determining alignments between questions and potential answers [67, 169, 168, 160, 189]. Over the years, the proliferation of Web and the abundance of text data triggered a growing interest in QA. In 1999, TREC added the seminal QA track that spurred development of practical QA systems. Traditionally, QA models were complex pipelines, composed of carefully crafted components that could be broadly categorized into: question processing, finding candidate answer documents, and answer selection [86]. Candidate answers were retrieved from a large-scale knowledge source that could be either a text collection or a Knowledge Base (KB). In 2011, IBM’s Watson DeepQA, devising a similar multi-faceted complex pipeline, marked a major milestone in QA by beating human champions in the trivia quiz show, Jeopardy!¹.

Retrieval has invariably been ingrained at the heart of QA. The community has come a long way, especially in the deep learning era and after the ubiquity of large

¹<https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html>

Pre-trained Language Models (PLM). Nowadays, modern QA models unequivocally rely on “retrieval”, although not necessarily analogous to traditional IR where actual documents are fetched [101, 70, 90, 192, 170]. Retrieval may be done as a kind of inductive bias during training [103, 21], or implicitly over a vast parameter space [153, 22, 32].

Notwithstanding the successful reign of modern neural models and the presence of plentiful resources, the task of QA still remains an open problem and a handful of major challenges yet to be tackled. Different questions may require different reasoning skills [34] for answering them because “the process of drawing conclusions” [156] to find an answer is not necessarily the same. Some questions require complex reasoning such as multi-hop reasoning [195], discrete reasoning [5, 47], coreferential reasoning [40], and spatial reasoning [133], whereas others do not. Unfortunately, the gap between the state-of-the-art models and human performance in many complex reasoning benchmarks highlights that there is room for improvement. Despite some recent efforts [93, 113], it is also not clear how to integrate the reasoning capabilities to build an intelligent unified QA system. Moreover, numerous studies have revealed the brittleness of the state-of-the-art QA models under a variety of realistic scenarios. QA models undesirably learn spurious patterns such as matching local textual patterns [82] or notoriously fail to generalize to new domains [130, 163].

1.1 Thesis Statement

In this thesis, I argue that **robust and scalable QA systems can be built using knowledge acquisition methods**. The novelty of this statement is embedded in the simultaneous attention to robustness and scalability. This dissertation aims at introducing principled practices towards this goal. To this end, I first explain the key aspects of the problem in this section.

1.1.1 Answering Information-seeking Questions

Questions can be framed in a myriad of ways, depending on the intent of the inquirer. In this dissertation, we focus on one particular type of questions, called *information-seeking questions*. This type of questions are typically asked when a person seeks information that they do not have [156] (*e.g.* *Who did Bette Midler portray in the Rose?* or *What’s the dog’s name on Tom and Jerry?*). Information-seeking questions are often acontextual, as opposed to questions in reading comprehension tasks where a question should be answered within a given context. In real-world, questions are not always perfectly formulated and may be rife with ambiguity [132] and presupposition [98], thus posing yet more challenges for answering such questions. A crucial challenge in QA is that users who write questions do not know the answer, and thus the likelihood of using question words that are likely to appear in answer documents diminishes, unlike in IR where users use words that are expected to be present in relevant documents. This problem is known as lexical chasm [20].

My main focus in this dissertation is on factoid questions whose answers are mostly short and extractive. Factoid questions often target specific factual statements, whereas non-factoid questions, usually starting with “why” or “how”, demand explanatory answers that are long and may span several paragraphs. Focusing on factoid questions allows us to account for the most common scenarios in knowledge-intensive user activities.

The task of automatically answering information-seeking questions, also referred to as open-domain question answering [180], involves sifting through a massive knowledge source. The knowledge source (*e.g.* CommonCrawl, and IR document collections such as ClueWeb or Robust) consists of documents, written in natural language text, spanning a wide variety of subjects.

Traditionally, IR models and search engines were part of a rather complex pipeline to narrow down the search space. With the rise of deep neural networks, open-domain

QA models are reduced to only two components: a traditional bag-of-words retriever along with a deep neural model that finds the answer given the retriever output [26]. However, bag-of-words models loudly fail when a match can only be found based on the meaning rather than the surface form. The dominance of Transformer-based LMs [43, 147] that are *pre*-trained on a colossal volume of unlabelled text, combined with transfer learning, has paved the way for effective dense retrievers [101, 90] that surpass traditional IR retrievers. More recently, retrieval-augmented models [103, 21, 11] that perform a non-parameteric nearest neighbours search over a massive scale memory are shown to work well for knowledge-intensive NLP tasks.

1.1.2 Knowledge Acquisition

Acquiring knowledge from unstructured text endows machines with problem-solving capabilities that is often warranted to tackle real-world language tasks. Machines broadly manifest acquired knowledge in various forms, ranging from explicit symbolic forms to implicit memories under the guise of weight matrices in neural networks. To tell if a system has actually acquired knowledge, one generally needs a verification mechanism. Specifically, the system is expected to lose its ability to solve the problem at hand, had the underlying knowledge not been presented to it.

In this dissertation, we leverage knowledge acquisition to tackle open-domain QA, a long-standing problem in natural language processing and information retrieval communities. The ability to successfully answer questions offers strong versatility to fulfill the information need of users since users can formulate nearly all their information need in the form of questions. More importantly, the proliferation of textual content expressing factual statements serves as a valuable resource that can be used to acquire knowledge from.

1.1.3 Robustness

In the deep learning era, QA models have achieved impressive performance in both reading comprehension (closed-domain) and open-domain QA. On some benchmarks like SQuAD [149, 148], QA models have actually surpassed human performance. Despite the substantial progress in QA, these models are repeatedly shown to be fitful [82, 130, 15] when the test distribution differs from the training distribution [97] that often arises in practice. The fragility of models under distribution shift impedes their deployment and limits their trustworthiness in real-world applications. We say a model is robust when its performance does not substantially degrade under distribution shifts. Robust models avoid shortcuts [63] and spurious patterns [126, 140], and instead, learn invariances across environments [9]. Two common causes of distribution shift are:

1. **Adversarial attacks** [66, 82, 15] where the input is synthetically perturbed to deceive the model into wrong predictions.
2. **Domain shift** where the test data is drawn from related but distinct domains, compared to the training set [97]. In this sense, the concept of *domain generalization* is synonymous with robustness under domain shift.

In this dissertation, my main focus is on *out-of-domain generalization* that is more likely to naturally occur in real-world scenarios. Here, a domain is a manifold that spans a variety of dimensions including genre, scope, stylistic conventions, etc. [156].

1.2 Key Contributions

The main claim of this dissertation revolves around building QA systems that acquire knowledge from unstructured text and are capable of using the knowledge to answer information-seeking questions. Our proposed techniques are designed to make QA systems safer to be used in the wild in that they are shown to be robust enough

under domain shift. Overall, the desiderata for building such QA systems can be summarized as follows:

- D1:** *Knowledge acquisition* from unstructured text, either explicitly or implicitly, because the underlying knowledge may serve as a kind of inductive bias or a prerequisite.
- D2:** *Robustness* when test distribution differs from training distribution because distribution shifts imminently occur in most real-world scenarios.
- D3:** *Scalability* in building models that can scale up to large collections, thereby fostering *efficiency* that saves compute and energy.
- D4:** *Reliable evaluation tools*, including annotated datasets and metrics, to accurately gauge the performance of systems and to offer ample diagnostic tools when issues arise.

I tackle this problem from four different perspectives: retrieval, QA machine reader, meta information, and evaluation. My contributions mainly encompass modelling, data-centricity, and evaluation; all constitute necessary steps for building systems that benefit real-world scenarios and do not easily break when data distribution shifts.

In summary, the key contributions of the present work are:

- **Local Word Order in Sparse Retrieval:** Traditional sparse retrieval has been popular in open-domain QA and is still considered as a strong baseline. In fact, a winning recipe is to combine sparse retrievers with dense retrievers. However, sparse retrievers ignore salient syntactic cues such as word order that offer useful hints for retrieval, especially in knowledge-intensive NLP tasks such as open-domain QA where localized context is crucial. To this end, I introduce a simple and yet, effective retrieval method that addresses mismatches, rooted in syntactical discrepancies (Chapter 2). My proposed model is contingent on the word order by leveraging n -grams in retrieval and is reminiscent of term dependence in IR models.

- **Data Augmentation and Robustness:** Data Augmentation (DA) is a tried-and-true technique to overcome the scarcity of training data. It is also well-documented that DA can make models more robust [72, 73, 151]. To this end, I present a novel taxonomy of DA techniques based on the training strategies (Chapter 3). Further, I provide an empirical analysis that corroborates the improvements in out-of-domain generalization using DA across a variety of language understanding tasks including reading comprehension (Chapter 3).
- **Sample Efficiency and Data Augmentation:** Despite its advantages, DA substantially slows down the training. As a remedy, I devise a general framework, inspired by adversarial training, that can accelerate training on augmented data and can be plugged into any DA method (Chapter 3). I empirically show that our sample-efficient DA method retains the effectiveness while substantially speeding up the training.
- **Leveraging Meta Information:** The content-rich textual data in Web is often accompanied by some auxiliary data, known as meta information (*e.g.* document timestamp and location tags in Wikipedia). This supplementary data, when present, offers useful signals that can aid QA models in resolving ambiguity or more generally, in pinpointing answers. However, meta information is not always present. Thus, the initial step is to find ways to generate viable meta information. To this end, I aim at generating spatial auxiliary data, and develop an unsupervised algorithm that grounds location mentions in text to their corresponding geographic footprints; the task is known as *toponym resolution* (Chapter 4). An unsupervised method is a reasonable choice here because of the lack of sufficiently large annotated data.
- **Document-level Reasoning Benchmark:** Open-domain QA models invariably perform retrieval at the passage level mainly because passage retrieval is able to capture localized contexts. However, I identify a small set of questions

in well-known open-domain QA benchmarks that are impossible to be answered using passage retrieval (Chapter 5). Answering these questions require a larger context because they are reliant on the document narrative or the document structure. Interestingly, these questions are overshadowed by passage retrieval-based questions in existing datasets. Thus, I curate a novel benchmark from widely adopted benchmarks in which document-level evidence is critical in answering questions.

1.3 Dissertation Layout

This dissertation is organized into 6 chapters. Each chapter can be viewed as a piece of a puzzle and covers one aspect of my ultimate goal — building robust knowledge acquisition techniques for open-domain QA. Putting all these pieces together shapes my recommended principles towards robust open-domain QA models. After the introduction in Chapter 1, the question of building a robust information-seeking QA system is contemplated in the succeeding chapters. Chapter 2 concerns retrieval models that are well suited for open-domain QA where compositionality of meaning has a profound role in retrieving relevant documents. Chapter 3 studies the role of DA and sample efficiency in model robustness and presents a taxonomy of DA methods. Chapter 4 introduces an unsupervised toponym resolution method as a step towards creating rich content that will further be helpful in QA. Chapter 5 highlights an unheeded challenge in existing open-domain QA benchmarks where document-level reasoning is often overlooked. Finally, Chapter 6 summarizes the contributions and discusses potential future research avenues.

Chapter 2

Revisiting Local Word Order in Retrieval

In this chapter, I present a simple model that equips sparse IR models with syntactical cues in text that are paramount in QA. Specifically, the proposed retrieval model fulfills two desiderata, specified in Section 1.2: knowledge acquisition (D1) in that it leverages local word order from text, and scalability (D3) in that sparse retrievers are easy-to-implement and scale well to large document collections.

2.1 Introduction

The task of open-domain Question Answering (QA) involves answering questions over a massive collection of documents. Classical IR models (*e.g.* BM25 [154]) have been particularly popular as a retriever in this task [26, 184, 33, 185, 194] thanks to their simplicity and their scalability. While suitable for modelling coarser levels of relevance [122], sparse retrievers are not designed for NLP tasks where understanding the intricacies of human-written text is paramount [69, 52]. In QA, in particular, matching questions with documents via only lexical information ignores salient syntactic cues such as word order that offer useful hints for retrieval. Such mismatches propagate undesirable noise to the reader, curtailing the overall effectiveness of retriever-reader pipelines.

The shortcomings of sparse retrievers heralded the emergence of dense retrieval

models. By performing matching in an embedding space, dense retrievers [101, 70, 90, 192, 146] are capable of finding documents that are semantically close to questions, thereby mitigating vocabulary mismatch. However, they still rely on sparse retrieval as a first-stage retriever to construct an initial candidate pool [184], or as a complementary retriever with which their retrieval output is combined [90]. Also, despite their advantages, dense retrievers are no panacea. Compressing a question/-passage into a low-dimensional space cannot accommodate all intricate information, conveyed in text [94], especially for entity-centric questions [13]. Moreover, if the distribution of the test data shifts from training data, their matching effectiveness deteriorates [104, 177].

In this chapter, we study the role of word order in tackling mismatches, rooted in syntactical discrepancies, to adapt sparse retrievers for QA. We hypothesize that there is a direct relationship between the ordering of words in an answer passage and that of the question and that this relationship provides hints that can improve the retrieval. This is in large part because word order organically carries some semantic weight, which is often transferred from questions to answer passages. We introduce a frustratingly simple retriever model, based on the well-known query likelihood framework [144], that is aware of the local word order. In particular, our model is built atop BM25 via: (1) treating *bigrams* as the smallest unit for local word order, and (2) adding bigrams as a single compound term to the vocabulary. We also prescribe pre-processing strategies to find *impactful* bigrams. Our experiments on two standard open-domain QA benchmarks reveal that local word order-aware retrieval consistently outperforms BM25 and when combined with a dense retriever improves upon several strong dense retrievers.

Our contributions can be summarized as follows:

1. We propose a local word order-aware retrieval model that can be plugged into BM25 or other sparse retrievers;

2. We provide analysis on how impactful bigrams can be selected in retrieval, analogous to the common practice of stopword removal in IR; and
3. We demonstrate that large performance gains can be achieved in open-domain QA pipelines when local word order is incorporated.

2.2 Related Work

Traditional retrieval-based pipelines. Early open-domain QA models often consist of several carefully designed stages [99, 48, 14], their components can be summarized as follows: (1) *question processing*: extracts useful information such as expected answer type, (2) *question reformulation*: rewrites questions into search queries, and (3) *search engine*: finds relevant documents, (4) *post-processing*: pinpoints the exact answer from the relevant documents.

Retriever-reader models. Traditional IR models are frequently used in open-domain QA pipelines thanks to their simplicity and efficiency. Several retriever-reader pipelines [26, 33, 194, 131, 185] use sparse retrieval models to retrieve documents. However, sparse retrieval models struggle when vocabulary mismatch occurs, which is often tackled via: (1) Document/Question expansion: doc2query [141], and GAR [123], (2) Re-ranking [184, 100, 138], (3) Dense retrieval using a two-tower *bi-encoder* architecture [162, 101, 90, 192, 146, 94, 59].

Massive-scale monolithic models. Recently, massive-scale language models—e.g., GPT-3 [22], and T5 [153]—have shown incredible generalization capabilities on downstream NLP tasks including QA.

Incorporating word order in IR Traditionally, IR systems may incorporate n -grams and phrases as a form of term dependency to model natural language [175, 167, 50]. Moreover, the dependency between n -grams and their co-occurring terms

is shown to be a useful signal to determine the importance of n -grams [166]. More recently, in neural IR, n -grams are projected to an embedding space, allowing for a semantic matching [38]. Term dependence may also be modelled by adding features related to contiguous term spans into the ranking function [129, 18, 39].

2.3 Local Word Order-Aware Retrieval

The bag-of-word model is at the center of sparse retrievals in open-domain QA, but this model overlooks the underlying word order in questions and answer passages. The ordering of words in natural language text carries some meaning, which is often transferred from questions to answer passages. For this purpose, we draw a connection between word order and the well-known query likelihood model [144] in IR. According to the query likelihood model, the probability of a document D is gauged via its relevance to a query Q , denoted as $p(D|Q)$, which is proportional to $p(Q|D)$ based on Bayes’ rule. To account for term dependence, Metzler and Croft expand $p(Q|D)$ over query constituents $C(Q)$:

$$p(Q|D) \stackrel{\text{rank}}{=} \sum_{c \in C(Q)} \lambda_c f(c) \quad (2.1)$$

BM25 can be written based on Eq. (2.1) where query terms form $C(Q)$ and $f(c)$ refers to the scoring function. In SDM [129], term dependence is modelled via relative frequency of contiguous text spans and thus, $C(Q)$ represents all text spans of a fixed-length within the query. Inspired by this formulation, we build a scoring function atop BM25 by two modifications: (1) we employ only bigrams—pairs of words (w_1, w_2) where w_2 immediately follows w_1 —because longer text spans become highly sparse; (2) treating bigrams as a single compound term that can be seamlessly integrated into any scoring mechanism, analogous to DrQA [26].

A blind integration of bigrams is less helpful due to dependency relationships between their constituent unigrams and a likely presence of *unnecessary* bigrams. We

need a strategy to detect such bigrams and discard them. This practice actually exists in IR where eliminating stopwords from the vocabulary is a standard routine.

Which bigrams are impactful? Intuitively, bigrams are expected to carry more information than their subsuming unigrams to improve retrieval. Specifically, we compute mutual information between a bigram b and its constituent unigrams (w_1, w_2) where the amount of information is gauged via IDF:

$$\begin{aligned}\psi(b) &= \max(\text{MI}(b; w_1), \text{MI}(b; w_2)) \\ &= \frac{\text{IDF}_b}{\max(\text{IDF}_{w_1}, \text{IDF}_{w_2})}\end{aligned}\tag{2.2}$$

The effect of ψ , namely *selectivity*, is anchored in the frequency of bigrams with respect to their most rare word. In particular, if the frequency of a bigram remains close to its unigrams, adding this bigram to scoring is not expected to bring improvements. We empirically found that a minimum ψ of 1.2 works best in our experiments.

Importance of stopwords Unlike ad hoc IR where stopwords are removed, retaining stopwords within bigrams interestingly reduces the ambiguity, hence improving the retriever effectiveness in open-domain QA. This is simply because stopwords inside bigrams represent localized contexts, which may be important in finding answer passages.

For example, for the TV show “*Who Wants to Be a Millionaire*,” no bigrams will be selected if stopwords are discarded. Nonetheless, the co-occurrence of the bigrams “*who wants*” and “*a millionaire*” are strong evidence to assign higher scores to relevant documents about the show than other documents.

Word order reversal The ordering relationship in a question is sometimes reversed when it is transferred to an answer passage because of the interrogative mood of questions. Since we seek to retain only the components that are expected to transfer

to answer passages, we transform questions into their corresponding declarative form before extracting n -grams. We use a rule-based tool, namely QA2D¹ [42], for this purpose. Because QA2D requires an answer to generate a sentence, we define a special token as the potential answer and pass it to the tool.

2.4 Experiments

2.4.1 Experimental Setup

Datasets: We adopt two well-known information-seeking QA benchmarks in which questions are written independent of answer passages. The details of these datasets are provided in Table 2.1.

- Natural Questions-OPEN (NQ-OPEN) [101]: Originally derived from Natural Questions (NQ) [98], this dataset serves as an established benchmark for factual question answering. NQ is curated from Google search queries for which answers can be found in Wikipedia. Questions in NQ are often written in spoken language and sometimes are framed in imperative mood (*e.g. list all the planet of the ape movies*). For open-domain QA, a subset of NQ whose answers are not longer than 5 tokens were selected [101].
- TriviaQA (TQA) [85]: TQA questions are written by trivia enthusiasts who tend to formulate long and well-formed questions. It is comprised of trivia questions mined from a variety of quiz-league websites.

Since the original test sets are hidden in both datasets, the original dev sets are used as an unseen test set [101, 90].

Evaluation Metrics: We report standard IR metrics, hits ratio (hits@ k) and mean reciprocal rank (MRR@ k), for retrieval. The end-to-end effectiveness of the pipeline is measured using exact-match (EM) accuracy and macro-averaged F1-score (F1) [149].

¹<https://github.com/kelvinguu/qanli>

Table 2.1: Statistics of QA benchmarks. $|Q|$ denotes the average number of words in questions.

Dataset	#train	#dev	#test	$ Q $
NQ-OPEN	79.2K	8.8K	3.6K	12.5
TQA	78.8K	8.8K	11.3K	20.2

Sparse retrieval: Our knowledge source is Wikipedia that is released in DPR² [90]. The knowledge source consists of roughly 3.2M Wikipedia articles, equated to 21M passages of 100 words, and is composed of 6.9M and 110.6M unique case-folded unigrams and bigrams. We implement our sparse retrievers using Pyserini [111].

2.4.2 Passage Ranking

We compare our retriever, BM25_{OURS}, with other prominent passage ranking models in open-domain QA. Our baselines are categorized into three groups: sparse, dense, and hybrid. In sparse retrieval, BM25 results are taken from DPR [90]. In addition, Ma *et al.* [120] replicated the experiments in DPR whose results are shown as BM25_{Repl.} and DPR_{Repl.}. For hybrid retrievers, we combine dense retrievers with our sparse retrievers, as done in DPR, where a hybrid score is calculated by linear interpolation: $\text{BM25}(q, p) + \lambda \cdot \text{sim}(q, p)$. We follow DPR [90] and set λ to 1.1 in our experiments.

Table 2.2 presents the passage ranking results. Among sparse retrievers, BM25_{OURS} outperforms unigram BM25_{Repl.} retrieval models by 2.1% and 1.0% in terms of hits@100 on NQ-open and TQA, respectively. Moreover, BM25_{OURS} outperforms sequential dependence model³ (SDM) [129] where term dependence is modelled via the frequency of contingent term spans by around +0.5% in terms of hits@100 on both datasets. When combined with DPR, our simple hybrid retriever outperforms GAR [123], the state-of-the-art retrieval model (*i.e.* +0.2% and +0.5% hits@100 gains on NQ-OPEN and TQA, respectively). Our results highlight that (1) local word order plays a pro-

²<https://github.com/facebookresearch/DPR>

³We used the implementation provided in the Anserini toolkit [111].

found role in sparse retrieval, and (2) combining a dense and a sparse retriever is an effective recipe for retrieval.

Table 2.2: Hits ratio at top- k passages for various retrievers.

Model		NQ-open		TQA	
		hits@20	hits@100	hits@20	hits@100
<i>Dense</i>	DPR [90]	79.4	86.0	78.8	84.7
	DPR _{Repl.} [120]	79.5	86.1	78.9	84.8
	ANCE [192]	82.1	87.9	80.3	85.2
	RocketQA [146]	82.7	88.5	-	-
	DPR+GAR [123]	81.6	88.9	82.1	86.6
<i>Sparse</i>	BM25 [90]	59.1	73.7	66.9	76.7
	BM25 _{Repl.} [120]	62.9	78.3	76.4	83.2
	SDM	60.2	74.4	68.2	79.6
	BM25 _{OURS}	66.1	80.4	78.7	84.2
<i>Hybrid</i>	DPR + BM25 [90]	78.0	83.9	79.9	84.4
	DPR + BM25 _{Repl.} [120]	82.7	88.1	82.6	86.5
	DPR + BM25 _{OURS}	83.2	89.1	83.5	87.1

2.4.3 End-to-End Results

We evaluate the effectiveness of the overall pipeline by plugging two well-known readers into our retrieval model: DPR [90] (*i.e.* DPR_{Multi}) as an extractive reader, and Fusion-In-Decoder (FiD) [81] as a generative reader. The readers here are based on BERT_{base}, so all models have roughly similar number of parameters. Following DPR, we retrieve the top-100 passages and feed them into the reader to obtain the final answer. We test the pipeline using two retrievers: BM25_{OURS}, and a hybrid retriever (*i.e.* DPR+BM25_{OURS} = Hybrid_{OURS}).

The results are reported in Table 2.3. Among extractive readers, our pipeline

with $\text{BM25}_{\text{OURS}}$ outperforms other baselines with sparse retrievers by around 4.3% and 0.9% on NQ and TQA, respectively. Furthermore, our hybrid retriever brings additional 2.8% gains on average and surpasses $\text{DPR}_{\text{Repl.}} + \text{Hybrid}$ [120] by around 1.8% margin on average. Similarly, for generative readers, we observe more than 2% gains on both datasets when using our hybrid retriever.

Table 2.3: Exact-Match (EM) accuracy of various open-domain QA pipelines.

	Model	NQ-open	TQA
<i>Extractive readers</i>	ORQA [101]	31.3	45.1
	DPR [90]	41.5	56.8
	DPR+BM25 [90]	32.6	52.4
	DPR+Hybrid [90]	38.8	57.9
	$\text{DPR}_{\text{Repl.}}$ [120]	42.5	58.3
	$\text{DPR}_{\text{Repl.}} + \text{BM25}$ [120]	37.0	59.2
	$\text{DPR}_{\text{Repl.}} + \text{Hybrid}$ [120]	43.2	60.0
	DPR+RocketQA [146]	42.8	-
	ANCE pipeline [192]	46.0	57.5
	DPR+GAR [123]	41.8	62.7
	DPR+ $\text{BM25}_{\text{OURS}}$	41.3	60.1
	DPR+ $\text{Hybrid}_{\text{OURS}}$	<u>44.9</u>	<u>61.9</u>
<i>Generative readers</i>	RAG [103]	44.5	56.8
	FiD_{Base} [81]	48.2	65.0
	$\text{FiD-KD}_{\text{Base}}$ [80]	<u>49.6</u>	68.8
	$\text{FiD}_{\text{Base}} + \text{Hybrid}_{\text{OURS}}$	50.4	<u>67.7</u>

2.4.4 Ablation Study

We study different versions of our model by including and excluding different classes of question unigrams u and bigrams b in BM25 :

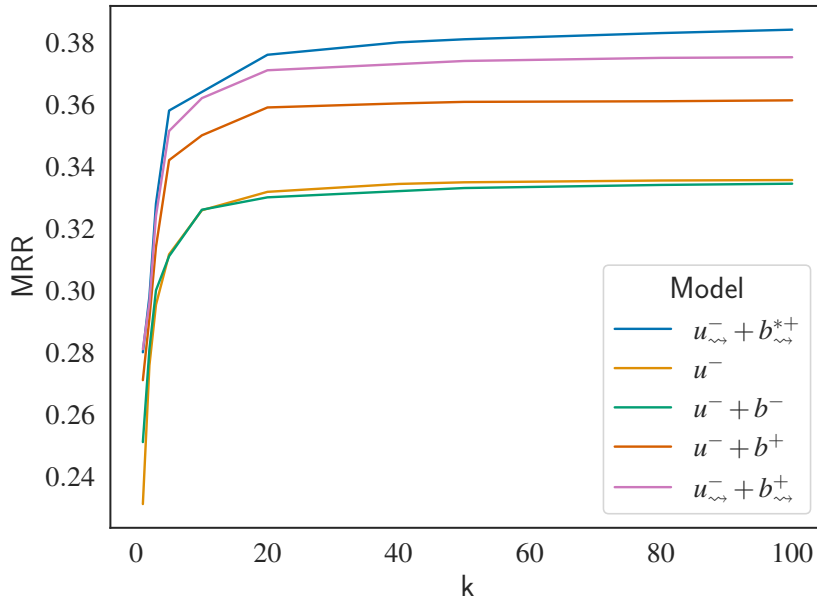


Figure 2.1: MRR@ k on NQ-open for 5 variants of our proposed models using BM25. $u_{\rightsquigarrow}^- + b_{\rightsquigarrow}^{*+}$ leads other models.

- (i) n -grams in which stopwords are absent ($-$),
- (ii) n -grams containing at most one stopword ($+$); for unigrams, this set is empty,
- (iii) Before extracting n -grams, questions are transformed into a declarative form, as explained in Section 2.3 (\rightsquigarrow),
- (iv) All bigrams except non-selective ones are considered, as discussed in Section 2.3 (b^*).

The results are showcased in Table 2.4. For unigram retrieval, we observe that dropping stopwords yields better performance, whereas applying question transformation leaves the results almost unchanged. Finally, by putting all pieces together, $u_{\rightsquigarrow}^- + b_{\rightsquigarrow}^{*+}$ on average yields 1.1% gains for BM25 in terms of hits@100, compared to unigram retrieval. Furthermore, we plot MRR@ k by varying k for several variants of our proposed retriever on NQ (a subset of models were selected to make the plot easier to read), as depicted in Figure 2.1. Similar to previous results, $u_{\rightsquigarrow}^- + b_{\rightsquigarrow}^{*+}$ easily surpasses all the variants.

Table 2.4: Hits ratio and MRR at top-100 passages for all variants of our sparse retriever. [†] and [‡] indicate statistical significance (p -value < 0.05) over u^- (row 2), unigram retrieval that is a widely-adopted retriever in open-domain QA, and over $u^- + b^-$ (row 5), a DrQA-inspired retriever, respectively.

	Model	Description	NQ-open		TQA	
			hits@100	MRR	hits@100	MRR
1	u	Unigrams	75.4	0.318	77.5	0.404
2	u^-	Unigrams excluding stopwords	78.6	0.336	83.8	0.568
3	u_{\rightsquigarrow}^-	Row 2 with question transformation	78.7	0.338	83.7	0.572
4	$u + b$	(Uni+Bi)-grams	78.1	0.328	80.5	0.436
5	$u^- + b^-$	(Uni+Bi)-grams excluding stopwords	78.6	0.334	83.9	0.573
6	$u^- + b^+$	Unigrams+Bigrams with ≤ 1 stopword	79.9 ^{†‡}	0.361 ^{†‡}	84.1	0.594 ^{†‡}
7	$u_{\rightsquigarrow}^- + b_{\rightsquigarrow}^+$	Row 6 with question transformation	80.4^{†‡}	0.375 ^{†‡}	84.1	0.597 ^{†‡}
8	$u^- + b^{*+}$	Row 6, but Bigrams chosen via Eq. (2.2)	79.8 ^{†‡}	0.366 ^{†‡}	84.0	0.590 ^{†‡}
9	$u_{\rightsquigarrow}^- + b_{\rightsquigarrow}^{*+}$	All together, BM25 _{OURS}	80.4^{†‡}	0.384^{†‡}	84.2	0.608^{†‡}

2.4.5 Document Ranking

In open-domain QA, passage retrieval is shown to be most effective [185, 194], but here, we conduct a document-level retrieval experiment to demonstrate the effectiveness of our proposed retriever at various granularity levels. At document-level, the retriever ranks Wikipedia articles. The effectiveness of retriever models are evaluated at top 10 documents—i.e., roughly equivalent to 100 passages. Based on our observation in passage ranking, we selected only effective models from the variants of our retriever. We compare the results with the following baselines:

- **DrQA⁴ [26]:** Analogous to TF-IDF with $u^- + b^-$, albeit with a difference that n -grams are mapped into 2^{24} bins using an unsigned murmur3 hash.
- **Extended DPR [90]:** We take the ordered list of passages, acquired by DPR

⁴<https://github.com/facebookresearch/DrQA>

and rank their subsuming documents, following [37]. More precisely, documents are ranked based on the maximum score of their passages (DPR-MaxP).

- **Hybrid:** Similar to passage ranking, the results of a dense retriever and a sparse retriever are consolidated by linearly interpolating their retrieval scores.

Our document ranking results, reported in Table 2.5, are consistent with passage ranking results. The full retriever—BM25 with $u_{\rightsquigarrow}^- + b_{\rightsquigarrow}^{*+}$ —achieves the best hits ratio and MRR among sparse retrievers and improves upon DPR-MaxP by a 1.9% and 0.8% margin on NQ-open, and TQA, respectively.

Table 2.5: Hits ratio and MRR at top-10 documents for various retrieval models. BM25 retrievers were run with different sets of unigrams and bigrams, explained in Table 2.4. † indicates statistical significance (*i.e.* McNemar’s test for hits@ k and Student’s t-test for MRR@ k with p -value < 0.05) over u^- .

Model	NQ-open		TQA	
	hits@10	MRR@10	hits@10	MRR@10
DrQA	71.8	0.511	80.2	0.663
BM25 with u^-	74.9	0.566	83.6	0.679
BM25 with $u^- + b^+$	78.5 [†]	0.592 [†]	84.4	0.692
BM25 with $u_{\rightsquigarrow}^- + b_{\rightsquigarrow}^+$	78.8 [†]	0.606 [†]	84.5	0.698 [†]
BM25 with $u_{\rightsquigarrow}^- + b_{\rightsquigarrow}^{*+}$	79.4 [†]	0.621 [†]	84.9 [†]	0.709 [†]
DPR-MaxP	82.3 [†]	0.667 [†]	86.2 [†]	0.731 [†]
DPR-MaxP+BM25 _{OURS}	84.2[†]	0.682[†]	86.8[†]	0.746[†]

2.5 Conclusions

Despite their simplicity, bag-of-word IR models often struggle in NLP tasks including open-domain QA where syntactic information offers plausible signals to find an answer. In this chapter, I showed that a simple modification to sparse retrievers to incorporate local word order has a substantial impact on their effectiveness. The

findings also revealed that although appealing, dense retrievers compromise their exact matching capabilities as their effectiveness easily improves when combined with our enhanced sparse retriever. I conjecture that this problem arises due to a lack of understanding over contiguous spans of text. Thus, building dense retrieval models that are designed to better represent such spans can be an interesting direction for future work.

Chapter 3

Robustness Via Sample-Efficient Data Augmentation

Data Augmentation (DA) is known to improve the generalizability of deep neural networks as well as making models more robust. This chapter describes a taxonomy over DA methods and points out that more effective DA methods tend to be prohibitively slow. Moreover, I present a framework to bridge the gap between efficiency and effectiveness in DA. Our method checks robustness (D2) and efficiency (D3) requirements from the desiderata, provided in Section 1.2.

3.1 Introduction

The undeniable importance of data in deep learning [157, 155] and the costly process of data annotation has propelled researchers into leveraging DA in a broad range of applications from computer vision [36, 183] to NLP including machine translation [161, 164], language understanding [164, 145, 46], and question answering [163, 4, 119, 12]. DA is shown to be effective in improving generalization of deep neural networks [44, 191] and in increasing the number of training samples especially in low resource data regimes [161, 199]. Nonetheless, in NLP, the discrete nature of text poses further complexity to DA as generating semantically viable text from another text is challenging [54].

DA methods can be broadly categorized into *task-aware* and *task-agnostic* meth-

ods. Task-agnostic DA methods essentially generate augmented text regardless of the task at hand and often do not warrant additional training or fine-tuning. They can be based on some hand-crafted heuristics [200, 186], back-translation [161, 49], or token replacement from a pre-trained language model [96, 190, 137]. Even though deploying task-agnostic methods is straightforward, these methods do not take into account any task-specific information, and thus, their performance is usually limited. On the other hand, task-aware DA methods are capable of generating augmented samples, conditioned on the downstream task objective [77, 191, 150]. These methods adapt augmented examples specifically for a task in that they construct augmented examples, sometimes partly, during training. Despite their advantages, they often incur additional training costs, resulting in a prohibitively slow and a computationally expensive training.

In general, the central problems surrounding DA techniques in NLP can be summarized as follows: First, DA methods are mostly not sample-efficient in that they add arbitrary number of augmented samples to the training data and naively incorporate all of them into training without investigating how many of augmented samples are actually needed. Second, although more effective, task-aware methods are notoriously time-consuming to train. This is especially problematic in large-scale datasets such as SQuAD [149] and MNLI [187]. Third, most DA methods are not universal as they work solely with a particular setup—e.g., training a single-network [191], or training in teacher-student settings [150]. Overall, the importance of both sample efficiency and training efficiency for DA has been often overlooked.

Motivated by the above problems, in this work, we introduce a universal DA method, Glitter ✨¹, which can be plugged into any DA method to make them sample-efficient, and task-aware without sacrificing performance. Specifically, given a pool of augmented samples that are generated offline, our proposed method follows a minimax approach [53] to select a small subset with maximal expected loss (*maximization*

¹Inspired by “*All that is gold does not glitter*” —J.R.R. Tolkien, *The Fellowship of the Ring*.

step) during training. Without any further adjustments to the training algorithm, the task objective can be optimized for this selected subset (*minimization step*).

Our key contributions can be summarized as follows:

1. Glitter ✨ is a universal method which can be effortlessly applied to any DA method to enforce sample efficiency while maintaining (or even boosting) their performance.
2. We devise strategies to adapt Glitter ✨ for a variety of widely used training setups including single-network, consistency training, self-distillation and knowledge distillation.
3. Through our empirical evaluations, we show that Glitter achieves superior performance over state-of-the-art DA methods on GLUE, SQuAD, and HellaSwag, while significantly speeding up the training.

3.2 Related Work

3.2.1 Task-agnostic DA in NLP

Contextual augmentation techniques [96, 190] use pre-trained language models for DA. Kobayashi [96] propose bidirectional LSTM language models for word substitution conditioned on the label of their input text. SSMBA [137] and TinyBERT [84] perturb the input by masking some of the tokens, and then, sample tokens from a BERT model to replace the masked tokens and generate augmented samples. Back-Translation [161] augments data using two consecutive translation models: the first model to translate the input into an arbitrary target language; then, a second model to translate the result back into its original language. Mixed-up [68] generates augmented samples based on interpolating word embedding and sentence embedding vectors. Shen *et al.* [164] introduce a set of cut-off techniques that zero out contiguous spans of the embedding matrix at token level, feature level and span level.

EDA [186] consists of simple word-level operations including synonym replacement, random deleting, random insertion and random swapping.

3.2.2 Task-aware DA in NLP

One approach to leverage task-specific information is to assign different weights to augmented samples based on their individual impacts on the model [196]. Although effective, the re-weighting mechanism largely ignores sample efficiency. Wu *et al.* [190] introduce a mask-and-reconstruct approach, namely c-BERT, that fine-tune a pre-trained BERT model to predict label-compatible tokens. CoDA [145] combines various label-preserving transformations with adversarial training jointly with a contrastive regularization objective. Unsupervised DA (UDA) [191] uses off-the-shelf DA methods and adds an auxiliary *consistency loss* to the training objective. However, UDA is not sample-efficient and it is designed only for a single-network setup; how to deploy it in other training scenarios such as knowledge distillation is not clear. Hu *et al.* [77] propose a reinforcement learning-based technique where the reward function is defined based on whether generated augmented samples are label-preserving or not.

3.2.3 DA for KD

KD [23, 74], initially proposed as a model compression technique, aims at transferring the knowledge of an already trained model, called *teacher*, to a smaller or a same-size *student* model. Several studies found that DA can significantly boost KD’s performance in NLP. TinyBERT [84] uses a task-agnostic DA technique for its task-specific fine-tuning. Kamaloo *et al.* [87] and Rashid *et al.* [150] showed that DA can also be tailored for KD. In particular, MATE-KD [150] tunes a separate masked language model in order to generate augmented samples with maximum divergence. Kamaloo *et al.* [87] and Du *et al.* [46] employ *k*NN retrieval to fetch augmented samples from a massive sentence bank.

Glitter differs from previous work in that it simultaneously focuses on sample

efficiency, and universality such that it can be freely used in any training setting.

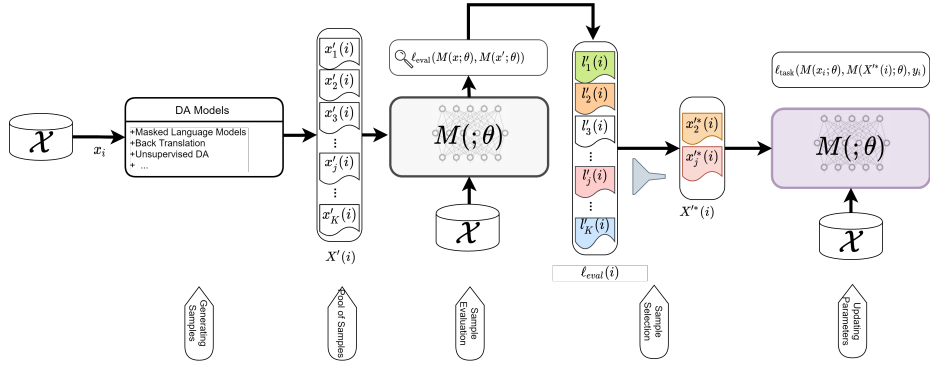


Figure 3.1: Illustration of Glitter (from left to right): first, generating augmented samples from different DA techniques; second, forming a pool of samples $X'(i)$; third, evaluating the augmented samples using the — loss; fourth, filtering the top- k_1 samples based on their corresponding —; fifth, updating the parameters of the model by minimizing the task loss $\ell_{\text{task}}(: \theta)$.

3.3 Methodology

In this section, we introduce our task-aware DA method, Glitter ✨, that aims at using an efficient number of augmented samples without sacrificing performance. Our proposed strategy is agnostic to DA methods; it can be seamlessly plugged into any DA method with any training setting to enforce sample efficiency.

Existing learning-based DA methods train a separate DA model and adapt its output for a particular objective function that is entirely task-dependent:

$$\begin{aligned} \phi^* &\leftarrow \min_{\phi} \ell_{DA}(M(\Omega(x; \phi); \theta)) \\ x'^* &= \Omega(x; \phi^*) \end{aligned} \tag{3.1}$$

where $\ell_{DA}()$ is a loss function, geared towards the objective of the task, $\Omega(; \phi)$ is the DA model with trainable parameters ϕ , and $M(; \theta)$ refers to the original model, parameterized by θ .

In contrast to learning-based DA, we propose to generate many augmented candidates using any arbitrary DA method prior training, and adaptively select most suitable candidates during training. This procedure does not introduce additional

trainable parameters into training, and more importantly, is capable of automatically ignoring unnecessary augmented examples. Let $(x_i, y_i)_{i=1}^N \in \{(\mathcal{X}, \mathcal{Y})\}$ represent training data such that a pair $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ contains an input example and its corresponding label. Suppose a pool of K augmented examples, $X'(i) = \{x'_k(i)\}_{k=1}^K$, are sampled from some DA model for each training example $(x_i, y_i) \in (\mathcal{X}, \mathcal{Y})$. Note that Glitter imposes no restrictions on how to augment training data; augmented samples can be generated via a single or even multiple DA models.

Sample Selection. Given a pool of augmented samples, our approach is to adaptively select the best candidates according to a particular defined criteria. Inspired by the minimax approach [53, 179], our selection mechanism is based on finding top- k_1 (out of K) worst-case augmented samples from the X' set. Minimizing the main model loss function on these worst-case augmented samples will help improving the generalization of the model [179]. In order to rank augmented samples, we evaluate $X'(i)$ based on a distance function with respect to the corresponding original training sample, x_i , within the model’s latent space:

$$\begin{aligned}
 X'^*(i) &\leftarrow \text{top}_{k_1} \left(\ell_{\text{eval}}(M(x_i; \theta), M(X'(i); \theta)) \right) \\
 X'^*(i) &= \{x'_j(i)\}_{j=1}^{k_1} \subset X'(i)
 \end{aligned}
 \tag{3.2}$$

where $\text{top}_{k_1}()$ denotes returns top- k_1 indices based on the scores returned by ℓ_{eval} , $X'^*(i)$ is the set of k_1 selected augmented samples for x_i ; $\ell_{\text{eval}}()$ is the evaluation loss which is determined via the task objective.

Updating the Model Parameters. After obtaining the top- k_1 augmented samples, we group them with the original training samples, $\{x_i\} \cup X'^*(i)$, and subsequently, update the model parameters only based on this selected set of augmented

samples on the original loss:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i=1}^N \ell_{\text{task}}\left(M(x_i; \theta), M(X'^*(i); \theta), y_i\right) \\ \theta_t &\leftarrow \theta_{t-1} - \lambda \nabla_{\theta}(\mathcal{L}(\theta))|_{\theta_{t-1}} \end{aligned} \tag{3.3}$$

where N is the number of training samples, λ is the learning rate, and $\ell_{\text{task}}()$ is the final task loss—e.g., cross entropy (ce) for classification—that is computed over both original data and selected augmented data. In the remainder of this section, we discuss how Glitter ✨ can be applied to popular training settings including general DA for single networks, and DA for teacher-student (KD) setups. Note that Glitter ✨ is not restricted to these settings and may be adapted for other settings such as DAIR [78].

3.3.1 General DA for Single Networks

We consider three potential setups for the single network scenario: (1) General single network, (2) Self-distillation, and (3) Consistency training.

General Single Network. In this setup, augmented samples are exploited in a semi-supervised manner where we can evaluate them based on the divergence of their predicted output $M(x'_k(i); \theta) = p(y|x'_k(i); \theta)$ from the ground-truth label or the prediction of the original corresponding training sample $M(x_i; \theta) = p(y|x_i; \theta)$ using the cross entropy loss, ℓ_{ce} :

$$\begin{aligned} \ell_{\text{eval}} &= \ell_{ce}(y_i, M(x'_k(i); \theta)) \\ \text{or} & \\ \ell_{\text{eval}} &= \ell_{ce}(M(x_i; \theta), M(x'_k(i); \theta)). \end{aligned} \tag{3.4}$$

The cross entropy criterion is not the only option here. Other choices for ℓ_{eval} include (but not limited to) focal loss [114], and tilted loss [105].

For the final task loss, ℓ_{task} we can deploy a standard cross entropy loss over both

training samples and their corresponding selected augmented samples:

$$\ell_{\text{task}} = \ell_{ce}(y_i, M(x_i; \theta)) + \frac{1}{k_1} \sum_{x \in X'^*(i)} \ell_{ce}(y_i, M(x; \theta)). \quad (3.5)$$

Consistency Training (CT) [191]. In this configuration, we can employ the same ℓ_{eval} introduced in Eq. (3.4). As a result, our method naturally selects top- k_1 most inconsistent augmented samples for each training sample. Then, the network is optimized to make predictions for input augmented samples that are consistent with predictions of their corresponding original training samples:

$$\ell_{\text{task}}^{\text{CT}} = \ell_{ce}(y_i, M(x_i; \theta_t)) + \frac{1}{k_1} \sum_{x \in X'^*(i)} \ell_{ce}(M(x_i; \theta_{t-1}), M(x; \theta_t)). \quad (3.6)$$

As stated in [191], the second term in Eq. (3.6) leverages the previous prediction of the network for each training example.

Self-Distillation (Self-KD). In Self-KD, we first train a model, and then, use it ($M(; \theta^*)$) as a teacher to train an identical model but initialized from scratch using KD [57]. How to adjust ℓ_{eval} and ℓ_{task} is detailed in Section 3.3.2.

3.3.2 DA for Teacher-Student (KD)

In this setup, we have a teacher model, $T(; \psi^*)$ with parameters ψ that is already trained on the training data, along with a student model, $M(; \theta)$, which we aim to train. The selection criterion for augmented samples is to maximize divergence between the teacher and the student:

$$\ell_{\text{eval}}^{\text{KD}} = \ell_{KL}(T(x'_k(i); \psi^*), M(x'_k(i); \theta)) \quad (3.7)$$

where ℓ_{KL} refers to the KL divergence. After selecting the maximum divergence augmented samples, then we calculate the KD loss as following:

$$\ell_{\text{task}}^{\text{KD}} = \alpha \ell_{ce}(y_i, M(x_i; \theta)) + (1 - \alpha) \times \frac{1}{k_1 + 1} \sum_{x \in \{x_i\} \cup X'^*(i)} \ell_{KL}(T(x; \psi^*), M(x; \theta)) \quad (3.8)$$

where α is a hyperparameter.

3.4 Experiments

3.4.1 Setup

To incorporate unlabelled augmented data into training, we adopt CT [191] and KD [74]. To this end, we conduct experiments under two settings:

- **Standalone** where we train a single model on the augmented data. In this setting, we seek to answer two questions: (1) How much is DA capable of improving the model generalization? (2) Does sample efficiency of Glitter hurt performance? For this purpose, we fine-tune RoBERTa_{base} [118] using CT and Self-KD on augmented data.
- **Distilled** where we distill DistilRoBERTa [158] (student) from RoBERTa_{Large} [118] (teacher) using the augmented data. Note that the teacher is already trained on the original data and DA comes into play only during distilling the student model. Our goal here is to investigate whether DA is an effective means in knowledge transfer to curb the capacity gap [30] between a large model and a small one.

In both settings, we take the best performing model on the development set and evaluate it on the test set (depicted by *Test*). Additionally, for the standalone model setting, we also report results on the development set when models are trained only for 5 epochs (depicted by *Dev*), similar to CoDA [145], to make a comparison with baselines. Our *Dev* results are an average of 10 runs with different seeds.

DA Methods

We leverage three widely used textual augmentation methods:

1. **EDA** [186]²: We randomly replace 5% of the tokens with their synonyms and randomly delete up to 10%.
2. **Back-Translation (BT)** [161]: We use fairseq [142] to translate sentences into German and then back into English. We do nucleus sampling [76] with $p = 0.9$ for both translations. We find that $p = 0.6$ works better on sentiment classification.
3. **Mask-and-Reconstruct (MR)** [137]: We randomly mask 15% of the tokens and construct a new sentence by sampling from a pre-trained BERT_{Large} for masked tokens. We adopt top- k sampling with $k = 20$ to select new tokens. For MNLI, we obtain better results with top-10 sampling.

For each augmentation method, we generate 12 augmented examples per training instance for all datasets, except for large datasets (*i.e.* MNLI, QQP, and SQuAD) where the number of augmented examples are 8 per train example.

Baselines

Because the two environments (*i.e.* standalone and distilled) are different in nature, we compare Glitter with different baselines for each environment. For both, Vanilla-DA that takes all augmented data into account without reservation is the first baseline.

The baselines for the standalone setting are: CoDA [145], MMEL [196], and HiddenCut [27]. And for distilled, we consider MATE-KD [150].

3.4.2 GLUE

The GLUE benchmark [182] is a well-known suite of nine³ tasks that aim at evaluating natural language understanding models. We present test results in the distilled mode in Table 3.1. Glitter consistently outperforms Vanilla-DA, while it is faster to train. Specifically, Glitter achieves parity with Vanilla-DA for EDA in terms of the overall

²<https://github.com/makcedward/nlpaug>

³We excluded WNLI since our DA methods are not designed for this task.

average score, while scoring +0.2% and +0.4% higher for BT and MR, respectively. We observe that only in few cases Vanilla-DA negligibly outperforms Glitter (*e.g.* on MRPC, and STS-B for BT). Nonetheless, Glitter $8x/1x^4$ trains 50% faster than Vanilla-DA $8x$ on average, and 30% faster for $8x/2x$. Also, Glitter surpasses MATE-KD by +0.2% in the overall score. Unlike Glitter, MATE-KD introduces additional parameters to the model during training and it trains drastically slower because it generates augmented examples on-the-fly. Moreover, Table 3.1 illustrates that MR yields the best test results across the three DA methods except for SST where BT leads to better results. Based on this observation, we report results on MR augmented data for all GLUE datasets except for SST in the remainder of our experiments.

Table 3.1: Test results of the distilled experiment on GLUE. The augmentation size is $8x$. **Bold** and underlined numbers indicate the best and the second best results across the DA methods.

Method	CoLA	SST	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	Avg.
	MCC	Acc	Acc/F ₁	P/S	Acc/F ₁	Acc	Acc	Acc	
RoB _{Large}	63.8	96.8	90.6	92.4	81.5	90.3/89.8	94.8	88.3	87.3
BERT _{Large} [43]	60.5	94.9	87.4	87.1	80.7	86.7/85.9	92.7	70.1	82.5
DistilRoB	55.2	93.9	85.9	86.0	80.3	84.0/83.1	90.6	73.6	81.1
KD	54.9	94.0	86.8	87.3	80.5	85.1/83.7	91.9	73.5	81.7
<i>Task-Aware DA</i>									
MATE-KD [150]	56.0	94.9	90.2	88.0	81.2	85.5/84.8	92.1	75.0	<u>82.8</u>
<i>EDA [186]</i>									
Vanilla-DA	55.5	94.8	87.6	86.1	80.7	85.3/84.7	92.0	72.8	81.8
Glitter ✨	54.5	95.1	87.5	86.5	80.4	85.4/84.8	92.1	73.2	81.8
	<i>2x</i>	<i>1x</i>	<i>2x</i>	<i>2x</i>	<i>2x</i>	<i>2x</i>	<i>2x</i>	<i>1x</i>	
<i>Back-Translation</i>									
Vanilla-DA	53.4	95.1	88.5	87.5	80.9	85.9/ 85.9	<u>92.2</u>	73.5	82.1
Glitter ✨	54.9	95.1	88.4	87.3	80.9	<u>86.2/85.3</u>	<u>92.2</u>	73.7	82.3
	<i>2x</i>	<i>1x</i>	<i>1x</i>	<i>2x</i>	<i>2x</i>	<i>2x</i>	<i>2x</i>	<i>2x</i>	
<i>Mask-and-reconstruct</i>									
Vanilla-DA	<u>58.8</u>	94.5	88.7	87.0	80.9	85.8/84.9	91.8	74.0	82.6
Glitter ✨	59.2	95.1	<u>89.2</u>	<u>87.6</u>	<u>81.0</u>	86.6 /84.8	92.4	<u>74.1</u>	83.0
	<i>1x</i>	<i>1x</i>	<i>2x</i>	<i>1x</i>	<i>2x</i>	<i>2x</i>	<i>2x</i>	<i>2x</i>	

For the standalone mode, Tables 3.2 and 3.3 present the results on test and dev,

⁴Our notation is K/k_1 where K denotes the augmentation size and k_1 is the selection size.

respectively. Similar to distilled, Glitter outperforms Vanilla-DA by +0.5% for both self-KD and CT. Self-KD yields better results than CT on all GLUE tasks except CoLA. CT falls short on most GLUE tasks, compared to no DA results—i.e., top-2 rows in Table 3.2. This is why, we only evaluated Glitter with self-KD on the dev data. Glitter achieves superior performance gains, compared to all three baselines on all datasets except QNLI. The key advantage of Glitter is that the training procedure remains intact.

Table 3.2: Test result of the standalone experiments on GLUE using RoBERTa_{base}.

Method	CoLA	SST	MRPC	STS-B	QQP	MNLI-m	QNLI	RTE	Avg.
	MCC	Acc	Acc/F ₁	P/S	Acc/F ₁	Acc	Acc	Acc	
RoBERTa _{base}	61.9	95.4	88.6	89.3	80.4	87.6	93.0	81.6	84.7
Self-KD	61.7	95.7	89.0	89.0	80.8	88.3	93.0	81.7	84.9
+ Vanilla-DA	61.5	96.1	88.9	89.7	81.0	88.0	92.9	81.1	84.9
	<i>8x</i>	<i>8x</i>	<i>8x</i>	<i>8x</i>	<i>8x</i>	<i>8x</i>	<i>8x</i>	<i>12x</i>	
+ Glitter ✨	62.5	96.0	89.8	89.5	81.1	88.1	93.5	82.3	85.4
	<i>8x/1x</i>	<i>8x/2x</i>	<i>8x/2x</i>	<i>8x/2x</i>	<i>8x/2x</i>	<i>8x/2x</i>	<i>8x/2x</i>	<i>12x/1x</i>	
CT + Vanilla-DA	59.4	95.6	89.0	85.8	80.3	82.5	92.0	80.2	83.1
	<i>8x</i>	<i>8x</i>	<i>8x</i>	<i>10x</i>	<i>8x</i>	<i>8x</i>	<i>8x</i>	<i>10x</i>	
CT + Glitter ✨	62.7	95.8	89.2	87.9	80.9	84.1	92.9	81.8	84.4
	<i>8x/1x</i>	<i>8x/1x</i>	<i>8x/1x</i>	<i>10x/1x</i>	<i>8x/2x</i>	<i>8x/2x</i>	<i>8x/2x</i>	<i>10x/1x</i>	

Out-of-Domain Generalization

We also evaluate Glitter on OOD datasets. To this end, we test our models, already trained on GLUE tasks, on OOD datasets whose data distribution differs from the original data. In particular, here are our selected OOD datasets:

- SST: IMDB [121], IMDB-Cont. [61], and IMDB-CAD [91], as done by Chen *et al.* [27]. Although both SST and IMDB datasets are collected on movie reviews, IMDB reviews tend to be substantially longer than SST sentences.
- STS-B: SICK [124], a semantic relatedness dataset, created from image and video captions. SICK and STS-B are collected on roughly identical domains, but from different sources.

Table 3.3: Dev results of the standalone experiment on GLUE using RoBERTa_{base}. For MMEL, the results are obtained from our implementation.

Method	SST	MRPC	MNLI-m	QNLI	RTE	Avg.
	Acc	F ₁	Acc	Acc	Acc	
RoB [118]	94.8	90.2	87.6	92.8	78.7	88.8
CoDA [145]	95.3	91.7	88.1	93.6	82.0	90.1
HiddenCut [27]	95.8	92.0	88.2	93.7	83.4	90.6
MMEL [196]	94.6 ± 0.8	91.9 ± 0.4	88.1 ± 0.1	93.2 ± 0.1	85.3 ± 1.0	90.6
RoB	94.3 ± 0.1	91.6 ± 0.5	87.7 ± 0.1	92.8 ± 0.2	84.5 ± 0.8	90.2
Self-KD	94.3 ± 0.2	91.5 ± 0.3	87.9 ± 0.1	92.9 ± 0.2	84.0 ± 0.6	90.1
+ Vanilla-DA	95.4 ± 0.5	92.0 ± 0.3	88.2 ± 0.1	93.4 ± 0.1	84.4 ± 0.7	90.7
+ Glitter ✨	95.7 ± 0.2	92.2 ± 0.5	88.2 ± 0.1	93.4 ± 0.1	85.6 ± 0.7	91.0

- QQP: PAWS_{QQP} [201], analogous to [27], and MQP [127], a medical question similarity dataset.
- MNLI: SciTail [95], collected from school-level science questions, and similar to [27], A-NLI [139], and HANS [126].
- RTE: HANS [126].

Table 3.4 showcases the OOD results for the distilled mode. Glitter outperforms Vanilla-DA in most cases, and is on par with it for nearly the rest. The only exceptions are MQP, and PAWS_{QQP} where Vanilla-DA outperforms Glitter by almost 1% on average. Also, all models do not generalize well to PAWS_{QQP} and A-NLI because their performance is below a majority-class performance. Moreover, a fine-tuned DistilRoBERTa achieves the best OOD performance on HANS, highlighting that DA is not actually helpful for OOD accuracy on HANS.

Table 3.5 and Table 3.6 report the OOD results for standalone models on test and dev settings, respectively. Glitter overwhelmingly outperforms all the baselines with a few exceptions. In the dev results, the fine-tuned model with no DA achieves the

best OOD generalization on IMDB, and SciTail, while HiddenCut scores the highest on A-NLI with a 1% margin. Similarly, in the test results, Glitter trails Self-KD with no DA on IMDB, IMDB-CAD, and SciTail.

Table 3.4: OOD results for the distilled mode. **Bold** numbers indicate the best result across DistilRoB models.

<i>Trained On</i> →	<i>SST</i>	<i>SST</i>	<i>STS</i>	<i>QQP</i>	<i>QQP</i>	<i>MNLI</i>	<i>MNLI</i>	<i>RTE</i>
Method	IMDb	IMDb-CAD	SICK	MQP	PAWS_{QQP}	SciTail	A-NLI	HANS
	Acc	Acc	P/S	Acc/F ₁	Acc	Acc	Acc	Acc
RoB _{Large}	93.7	94.0	84.3	71.6	43.6	82.0	45.9	81.8
DistilRoB	90.2	92.5	79.6	67.3	36.3	74.8	27.8	71.3
KD	90.6	93.2	79.9	65.6	33.1	77.3	28.9	70.6
<i>EDA</i> [186]								
Vanilla-DA	91.8	92.9	80.0	59.9	38.0	75.8	27.3	66.6
Glitter ✨	91.2	94.0	80.0	64.0	36.6	75.6	28.8	65.6
<i>Back-Translation</i>								
Vanilla-DA	92.2	92.1	80.3	69.6	35.0	76.5	27.9	68.0
Glitter ✨	92.4	92.8	81.2	68.7	35.2	77.6	30.4	70.5
<i>Masked-and-reconstruct</i>								
Vanilla-DA	91.8	92.9	80.4	68.5	33.7	77.4	28.5	69.3
Glitter ✨	92.0	92.5	80.7	68.8	35.3	78.2	29.9	70.9

Table 3.5: OOD results for *test* settings in the standalone mode. **Bold** numbers indicate the best result.

<i>Trained On</i> →	<i>SST</i>	<i>SST</i>	<i>STS</i>	<i>QQP</i>	<i>QQP</i>	<i>MNLI</i>	<i>MNLI</i>	<i>RTE</i>
Method	IMDb	IMDb-CAD	SICK	MQP	PAWS_{QQP}	SciTail	A-NLI	HANS
	Acc	Acc	P/S	Acc/F ₁	Acc	Acc	Acc	Acc
RoB _{Base}	92.2	94.3	80.6	70.7	38.6	78.5	31.4	78.5
Self-KD	92.6	95.0	80.2	70.9	37.6	79.4	32.1	79.5
+ Vanilla-DA	91.8	94.8	81.5	71.4	38.8	78.4	31.5	79.3
+ Glitter ✨	92.0	94.8	81.7	72.1	39.4	79.1	32.7	80.1
CT + Vanilla-DA	90.6	92.1	76.6	70.6	38.3	76.6	30.3	78.4
CT + Glitter ✨	92.2	93.7	79.4	70.7	38.8	77.0	31.6	80.2

3.4.3 HellaSwag

HellaSwag [198] is a dataset for situated commonsense reasoning that involves picking the best ending given a context. We augment contexts in HellaSwag using only BT

Table 3.6: OOD results for *dev* settings in the standalone mode. For MMEL, the results are obtained from our implementation. **Bold** numbers indicate the best result.

<i>Trained On</i> →	<i>SST</i>	<i>SST</i>	<i>SST</i>	<i>MNLI</i>	<i>MNLI</i>	<i>RTE</i>
Method	IMDb	IMDb-Con.	IMDb-CAD	A-NLI	HANS	HANS
	Acc	Acc	Acc	Acc	Acc	Acc
RoB _{Base}	91.9 ± 0.3	90.0 ± 0.4	94.1 ± 0.4	31.0 ± 0.6	73.7 ± 0.7	78.3 ± 0.4
HiddenCut [27]	-	87.8	90.4	32.8	71.2*	-
MMEL [196]	91.6 ± 0.1	90.5 ± 0.7	94.5 ± 0.4	31.4 ± 0.6	74.5 ± 0.6	78.3 ± 0.3
Self-KD	91.9 ± 0.3	90.3 ± 0.5	94.4 ± 0.4	30.9 ± 0.4	73.5 ± 0.7	78.2 ± 0.4
+ Vanilla-DA	91.6 ± 0.4	90.2 ± 0.4	94.3 ± 0.3	31.3 ± 0.5	73.9 ± 0.4	77.8 ± 0.3
+ Glitter ✨	91.7 ± 0.2	90.6 ± 0.2	94.8 ± 0.2	31.8 ± 0.4	74.6 ± 0.3	78.4 ± 0.2

to ensure that the choices remain meaningful for the augmented contexts. Because our standalone results have been consistent with the distilled results, we report our results only in the distilled mode. According to our results demonstrated in Table 3.7, Glitter comfortably surpasses Vanilla-DA by a +2.3% margin.

Table 3.7: Dev results of the distilled experiment on two downstream tasks.

Method	SQuAD	HellaSwag
	EM/F ₁	Acc
RoB _{Large}	88.9/94.6	85.2
DistilRoB	80.9/87.9	42.9
KD	81.1/88.2	42.5
+ Vanilla-DA (8x)	81.8/89.1	41.8
+ Glitter ✨ (8x/2x)	83.6/90.3	44.1

3.4.4 SQuAD

SQuAD [149] is a crowd-sourced reading comprehension benchmark that consists of more than 100K questions, derived from Wikipedia passages. The task objective is to extract an answer span from a given question/passage pair. We augment questions in SQuAD v1.1 using only BT to ensure that the answer can still be found in the given passage for the augmented questions. Analogous to HellaSwag, we report our results

only in the distilled mode. As shown in Table 3.7, Glitter outperforms Vanilla-DA by +1.8% in exact-match accuracy on the development set.

We also evaluate our trained models under distribution shift by testing them on QA datasets from four different domains: Wikipedia, New York Times, Reddit, and Amazon product reviews [130]. The OOD results are presented in Table 3.8. Glitter is consistently superior to Vanilla-DA in all four domains.

Table 3.8: OOD results for models trained on SQuAD and tested on QA datasets from four different domains [130].

Method	Wiki	NYTimes	Reddit	Amazon
	EM	EM	EM	EM
RoBERTa _{Large}	84.4	85.9	76.6	74.4
DistilRoBERTa	76.6	78.1	66.2	62.9
KD	76.5	78.7	65.7	63.0
+ Vanilla-DA (8x)	77.3	79.0	65.9	63.3
+ Glitter ✨ (8x/2x)	79.3	80.7	68.1	64.7

3.5 Ablation Study and Discussion

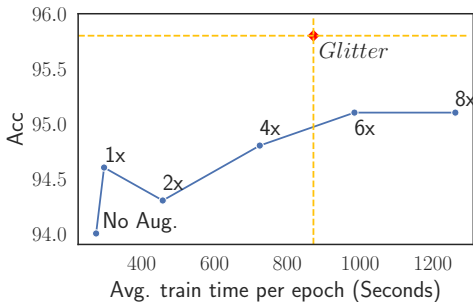
In this section, we aim to answer the following questions:

- How does training time of Glitter compare against Vanilla-DA?
- Instead of adaptively selecting augmented data during training, can we pre-process them to dispense with unnecessary examples prior to training?
- How many augmented examples are required for Glitter to work?
- Is our selection strategy based on sorting of ℓ_{eval} in Glitter important?

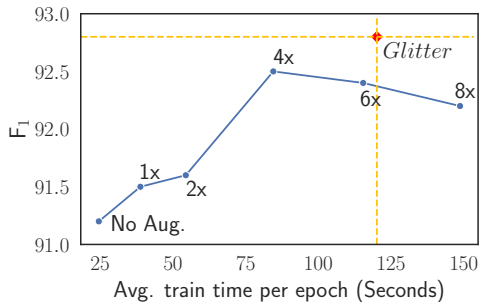
For this purpose, we conduct a detailed analysis on 4 GLUE tasks (*i.e.* SST, MRPC, QNLI, and RTE). We trained models based on Vanilla-DA and Glitter using Self-KD and tested them on the development set (the dev setting).

3.5.1 Runtime Analysis

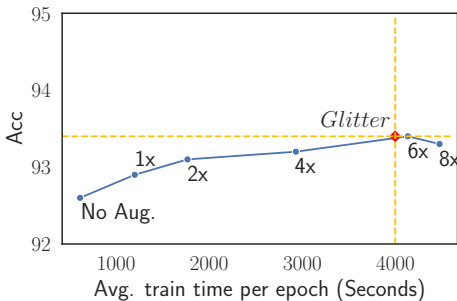
Throughout our experiments in Section 3.4, we compare Glitter with Vanilla-DA when number of augmentations are similar for both methods—i.e., $8x$. A natural question is: how would both DA methods behave with fewer augmented data? To this end, we vary augmentation size from $1x$ to $8x$ and train different Vanilla-DA models on each augmented dataset. We measure average the training time per epoch for all models. Figure 3.2 illustrates the dev accuracy as the training time increases. The training speed of Glitter $8x/2x$ is slightly faster than Vanilla-DA $6x$ on SST, MRPC, and QNLI and for Glitter $8x/1x$, is faster than Vanilla-DA $4x$ on RTE. Glitter is superior of the two on all datasets.



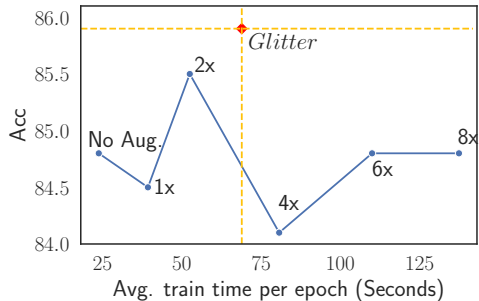
(a) SST



(b) MRPC



(c) QNLI



(d) RTE

Figure 3.2: Runtime Analysis of DA when training $\text{RoBERTa}_{\text{base}}$ using self-KD. The red point signifies Glitter $8x/2x$, except for RTE that is $8x/1x$.

3.5.2 Effect of Pre-processing Augmented Data

We conjecture that Glitter does not need any data engineering on augmented examples to obtain preferable performance gains. However, Vanilla-DA may require some pre-processing by weeding out potentially noisy data to become more effective. To investigate this, we exploit two pre-processing techniques:

- (1) **Confidence-based filtering:** Augmented examples for which the model’s confidence is below a minimum threshold β are discarded,
- (2) **Label-preserving augmentation (LP):** Augmented examples for which the model predicts a different label than the original example are discarded.

The results, reported in Table 3.9, show no meaningful performance gains by these pre-processing techniques. For Vanilla-DA, minimum confidence threshold of 0.7 performs slightly better as it brings minor improvements on MRPC (+0.3%) and QNLI (+0.1%), but is still lower than Glitter. On the other hand, applying these techniques slightly deteriorates the performance of Glitter in almost all cases. The only improvements are +0.1% on QNLI for LP and $\beta=0.7$.

Effect of Augmentation Size in Glitter. We explore how augmentation size affects the performance of Glitter. Throughout our experiments, we fix the augmentation size to $8x$, but now, we reduce augmentation size K to $6x$ and $4x$, while retaining selection size k_1 as before (*i.e.* 1 for RTE, and 2 for the rest). Our results, shown in Table 3.10, reveal that when K becomes close to k_1 , Glitter’s performance declines. Nonetheless, for a sufficiently large augmentation, Glitter starts to shine. For SST, and MRPC, the magic number is $8x$, whereas for QNLI, and RTE, Glitter performs best on $6x$. Another parameter in Glitter is the selection size k_1 . We find that for all tasks, the best value can be chosen from $\{1, 2\}$ (2 by default). Using this method, tuning k_1 is straightforward and does not impose additional complexity to our method.

Table 3.9: Dev results of self-KD exhibiting the effectiveness of different pre-processing techniques to filter augmented examples on 4 GLUE tasks. β and LP depict a minimum confidence threshold, and label preserving, respectively.

Method	SST	MRPC	QNLI	RTE
	Acc	F ₁	Acc	Acc
Vanilla-DA	95.1	92.2	93.3	84.8
+ $\beta = 0.7$	95.1	92.5	93.4	84.8
+ $\beta = 0.9$	95.0	92.2	93.3	83.8
+ LP	94.8	92.4	93.3	84.8
Glitter ✨	95.8	92.8	93.4	85.9
+ $\beta = 0.7$	95.0	91.5	93.5	85.2
+ $\beta = 0.9$	95.0	92.5	93.3	84.1
+ LP	95.1	92.2	93.5	85.9

Effect of Selection Strategy in Glitter. In this section, our objective is to assess whether our proposed selection algorithm is crucial in Glitter. To this end, we sample random augmented examples at each iteration, namely *Glitter-Rnd*, instead of selecting worst-case examples. As illustrated in Table 3.10 (the bottom two rows), the performance drops on all datasets—i.e., 0.2% on QNLI, and more than 1% on the rest, confirming the effectiveness of our selection algorithm.

3.6 Reproducibility

3.6.1 Fine-tuning Details

We adopted the publicly available pre-trained RoBERTa [118] and DistilRoBERTa [158] using the Huggingface Transformers library [188] and the Pytorch Lightning library⁵.

For the *test* settings, the model is evaluated on the development data once per epoch for small datasets and twice per epoch for large ones (*i.e.* SST-2, MNLI, QNLI,

⁵<https://github.com/PyTorchLightning/pytorch-lightning>

Table 3.10: Dev results of self-KD for studying the effect of augmentation size and the selection algorithm for 4 GLUE tasks.

Method	SST	MRPC	QNLI	RTE
	Acc	F ₁	Acc	Acc
	$k_1 = 2x$	$k_1 = 2x$	$k_1 = 2x$	$k_1 = 1x$
Glitter ✨ ($8x/k_1$)	95.8	92.8	93.4	85.9
Glitter ✨ ($6x/k_1$)	94.7	92.7	93.7	86.3
Glitter ✨ ($4x/k_1$)	95.0	92.1	93.3	85.7
Glitter-Rnd ($8x/2x$)	94.3	91.4	93.2	85.2
Glitter-Rnd ($8x/1x$)	94.3	91.8	93.2	84.5

SQuAD, and HellaSwag). The best performing model is chosen for testing. Our learning rate schedule follows a linear decay scheduler with a warm-up, specified as a ratio of the total number of training steps. Maximum number of epochs is set to 20 for all tasks except SQuAD, following [135]. For large datasets, we early stop with a patience of 10. The learning rate, and the batch size are tuned for each task separately. The details of hyperparameters are summarized in Table 3.12. We ran RoBERTa_{base} experiments with the similar hyperparameters, but with these exceptions: On QNLI, learning rate, batch size, and weight decay are set to 3e-5, 64, and 0.1; warmup ratio is set to 0.06 on QQP.

For *dev* experiments, we follow CoDA [145] on the GLUE tasks. Specifically, we train the model for 5 epochs with a batch size of 32, learning rate 1e-5, warmup ratio 0.06, weight decay 0.1, and linear learning rate decay. For SQuAD, and HellaSwag, the hyperparameters are detailed in Table 3.11.

All experiments were conducted on two Nvidia Tesla V100 GPUs.

3.6.2 Distillation Details

We implemented KD by caching the teacher’s logits prior to training. We performed grid search to find the best softmax temperature τ from {5.0, 10.0, 12.0, 20.0,

Table 3.11: Hyperparameters of DistilRoBERTa on two downstream tasks.

Hyperparameter	SQuAD	HellaSwag
Learning rate	1.5e-5	1.5e-5
Batch size	16	32
Max length	512	512
Max epochs	3	20
Warmup ratio	0.06	0.06
Grad. accumulation steps	4	1
Weight Decay	0.01	0.01
Softmax temp. τ (for KD)	5.0	10.0

Table 3.12: Hyperparameters of DistilRoBERTa on the GLUE benchmark. We used the same configuration for RoBERTa_{base} albeit with a few exceptions marked by (*).

Hyperparam.	CoLA	SST	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE
Learning rate	1e-5	1e-5	1e-5	1e-5	1e-5	3e-5/1e-5	5e-5*	1e-5
Batch size	32	64	16	32	64	64	128*	32
Max length	128	256	128	128	256	256	256	256
Warmup ratio	0.1	0.06	0.06	0.06	0.1*	0.08/0.06	0.08	0.06
Gradient acc. steps	1	4	1	1	4	4	4	1
Weight Decay	0.1	0.1	0.1	0.1	0.1	0.0/0.1	0.0*	0.1
Softmax temp. τ (for KD)	30.0	20.0	12.0*	12.0	20.0	12.0	12.0	12.0

30.0}. The value of τ used in our experiments are reported in Tables 3.11 and 3.12 for DistilRoBERTa and RoBERTa_{base}; with the exception $\tau = 20.0$ on MRPC for RoBERTa_{base}. Loss weight α , in Eq. (3.8), is set to 0.5 for all tasks except CoLA in which $\alpha = 0.75$.

3.7 Conclusion

In this chapter, I proposed a universal DA technique, namely *Glitter*, that can be freely applied to any DA technique to enforce sample efficiency without introducing additional parameters or changing the training procedure. I extensively evaluated

Glitter ✨ on a broad range of NLU tasks and in various widely used settings including consistency training, self-distillation and knowledge distillation and demonstrated substantial efficiency gains without compromising effectiveness. Extending Glitter to auto-regressive models for machine translation and abstractive summarization is an interesting direction for future work.

Chapter 4

An Unsupervised Model for Grounding Location Mentions

The Web data is often accompanied by attachments that offer supplementary information. For instances, images are geotagged by default when taken by smartphones, or news articles include authors, and datelines, or scientific articles provide an abstract along with some related keywords. Meta information provide useful means that can be used for data processing (*e.g.* data cleaning or data augmentation) or even modelling (*e.g.* inferring inductive biases). In QA, leveraging meta information has been successful [203, 116]. More recently, Asai *et al.* [13] constructed reasoning chains required for answering complex questions via Wikipedia link graph. In light of these work, QA systems, especially in real-world scenarios, can benefit from meta information. For instance, ambiguous questions may be successfully answered using some auxiliary information from the user. In this chapter, I take the very first step towards building such systems. More specifically, I present an algorithm that enriches text with auxiliary information, yet another step towards effective knowledge acquisition (D1) indicated in Section 1.2. The main focus here is on spatial information (*i.e.* geographic footprints). Hopefully, using methods such as the present work spur the development of QA systems leveraging meta information.

4.1 Introduction

The size of the Web has been growing near-exponentially over the past decade with a vast number of websites emerging on a variety of subjects and large volumes of textual data being made available every day. In particular, a staggering amount of Web content (such as news articles, blog, forum posts, and tweets) that are added online on a minute by minute basis make frequent use of location names as points of reference. However, many place names have multiple interpretations and using them as references introduces ambiguity which in turn leads to uncertainty. Determining geographic interpretations for mentions of place names, known as *toponyms*, involves resolving multiple types of ambiguities. Toponym resolution is the task of disambiguating or resolving toponyms in natural language contexts to geographic locations (i.e., the corresponding lat/long values). One of the formidable challenges is therefore related to resolving the ambiguity of place names. For example, consider the word *Paris* in the following sentences:

1. “The November 2015 *Paris* attacks were the deadliest in the country since World War II.”¹
2. “*Paris* was voted ‘the Prettiest Little Town in Canada’ by Harrowsmith Magazine.”²

The first sentence cites the tragic incidents in *Paris, France* while in the second sentence, the co-occurrence of *Canada* and *Paris* helps us identify *Paris*. This example illustrates that a toponym resolution method should probe for such clues in documents to reduce the inherent ambiguities of the natural language text. GeoNames³, the largest crowd-sourced location database, lists 97 interpretations for the place name *Paris*.

¹https://en.wikipedia.org/wiki/November_2015_Paris_attacks

²<http://www.brant.ca/en/discover-brant/paris.asp>

³<http://geonames.org/>

The problem of toponym disambiguation has been studied in the literature. Early works on geotagging documents rely on hand-crafted rules and heuristics (e.g., Web-a-Where [6]). Recent studies, however, are grounded on supervised and unsupervised models that do not warrant any manual rules [173, 128, 1, 41, 109]. Adaptive Context Features (or *Adaptive* in short), proposed by Lieberman and Samet [109], and TopoCluster, suggested by DeLozier *et al.* [41], are among the prominent methods that have been proposed in this area. Adaptive method casts toponym resolution as a classification problem, whereas TopoCluster leverages geographical measures to estimate geographical profiles for words.

In this work, we propose an unsupervised model to tackle toponym resolution since supervised methods yield a poor performance due to the paucity of sufficient annotated data. Our methods rely merely on the document content and a gazetteer primarily because supplementary information about a Web document often is neither available nor reliable. Clearly, any additional data such as the hosting site of the document and its location (if available) can further improve the performance.

Our toponym resolution model utilizes context-related features of documents. First, we develop a probabilistic model, called *Context-Bound Hypotheses (CBH)*, inspired by the work of Yu and Rafiei [197], to incorporate two context-related hypotheses into toponym resolution. Yu and Rafiei’s model aims at geotagging non-location entities and employs a primitive disambiguation technique to spatially resolve toponyms. We extend this model by integrating geographical information of locations into the hypotheses. These context-related premises capture some of the implicit relationships that hold between place names mentioned in the same document; thus, each toponym follows either the location of a frequent toponym or a nearby toponym. Then, we develop another model, called *Spatial-Hierarchy Sets (SHS)*, which discovers a minimal set of relationships (as discussed in Section 4.2) that can exist among toponyms. SHS maps the minimality problem to a conflict-free set cover problem wherein sets are constructed using containment and sibling relationships among toponyms. The final

model, *Context-Hierarchy Fusion (CHF)*, merges CBH and SHS to exploit context features in extracting minimal relationships.

We conduct extensive experiments to evaluate our model. Our experiments are carried out on multiple datasets, one collected and annotated by us and two others well-known and used in the literature, covering a large range of news sources. We assess the performance of our model and compare it with the state-of-the-art supervised and unsupervised techniques as well as a few commercial geotagging products including Yahoo! YQL Placemaker⁴, Thomson Reuter’s OpenCalais⁵, and Google Cloud Natural Language API⁶. Moreover, we study the generalization problem of supervised methods by feeding unseen data to Adaptive classifier, showing that the classifier cannot keep up with our unsupervised model.

In summary, the key contributions of this work are as follows:

- We devise an unsupervised toponym resolution model that leverages context features of documents as well as spatial relationships of toponyms to produce a coherent resolution.
- We extensively evaluate our model on different datasets and in comparison with state-of-the-art methods.
- We demonstrate that our unsupervised model surpasses the state-of-the-art unsupervised technique, TopoCluster [41], and that it can handle unknown data better than supervised techniques.

4.2 Problem Definition

Given a document \mathcal{D} and a sequence of toponyms $T = t_1, t_2, \dots, t_K$ mentioned in \mathcal{D} (*e.g.* extracted using a named-entity recognizer), toponym resolution refers to grounding each toponym t_i to a geographic footprint ℓ_i with a latitude and a longitude.

⁴<https://developer.yahoo.com/yql/>

⁵<http://www.opencalais.com/>

⁶<https://cloud.google.com/natural-language/>

Geographic footprints or references are often derived from a gazetteer, a repository of georeferenced locations and their associated metadata such as type/class, population, spatial hierarchy, etc. Following previous works [109, 173], we select GeoNames as our gazetteer primarily because not only is it the largest public location database with sufficiently high accuracy [2], but it also stores the spatial hierarchy of locations⁷. Additionally, the bounding boxes of some locations can be retrieved from GeoNames.

Each toponym t_i in \mathcal{D} has a set of location interpretations $L_i = \{l_{i,1}, l_{i,2}, \dots, l_{i,n_i}\}$, derived from a gazetteer \mathcal{G} , where n_i indicates the number of interpretations for toponym t_i . Hence, toponym resolution can be seen as detecting a mapping from location mentions T to location interpretations. The resolution method yet cannot enumerate all possible combinations of interpretations. For instance, in a document that contains only 6 U.S. states: *Washington* ($n_1=113$), *Florida* ($n_2=228$), *California* ($n_3=225$), *Colorado* ($n_4=230$), *Arizona* ($n_5=63$) and *Texas* ($n_6=53$), the number of possible interpretations exceeds 4 billion. The past works in this area therefore incorporate heuristics to reduce the immense search space. For instance, picking the most populated interpretation is a simple heuristic that has been adopted in early works [102]. However, population alone cannot be effective for an off-the-shelf resolution system. We address this problem by looking into containment and sibling relationships among toponyms in a document.

4.3 The Unsupervised Model

The proposed method leverages a combination of context-related features of documents to address toponym resolution. These features are grounded on the characteristics of toponyms. It is well-accepted (*e.g.* SPIDER [173]) that toponyms mentioned in a document often show the following minimality properties:

- *one-sense-per-referent*: all of the occurrences of a toponym generally refer to a

⁷OpenStreetMap, another well-known crowd-sourced gazetteer, is ruled out since it does not contain spatial hierarchies [65].

unique location within a single document;

- *spatial-minimality*: toponyms mentioned in a text tend to be in a spatial proximity of each other.

In this section, we develop context-bound hypotheses, inspired by the named entity geotagging method suggested by Yu and Rafiei [197]. Then, we describe spatial hierarchies built from containment and sibling relationships among location mentions in text. Lastly, we explain how these two methods coalesce into an unsupervised model to disambiguate toponyms.

4.3.1 Context-Bound Hypotheses

Yu and Rafiei [197] propose a probabilistic model to associate named entities to locations. The task of geotagging named entities is delineated as follows: given a named entity and a set of documents germane to it, a geotagger finds the geographic focus of the named entity. The model, introduced by Yu and Rafiei [197], incorporates two hypotheses: *geo-centre inheritance hypothesis* and *near-location hypothesis* and estimates the probabilities that these premises hold. The probabilistic model makes use of the known entities that are mentioned in the surrounding text to determine the geo-centre of a named entity. Their geotagging task mainly focuses on non-location named entities and does only a simple location disambiguation on each toponym, independent of other toponyms in the same document. A question here is if their probabilistic model can be applied to toponym resolution. This is the question we study in our *Context-Bound Hypotheses (CBH)* model. In particular, to model the cohesion of toponyms to context, we integrate the hypotheses with geographical information of locations in order to spatially locate a place mention. Context-Bound assumptions allow us to reduce toponym resolution to a probabilistic model, which we are set to compute the estimations in this section.

The space of possible interpretations (as shown with an example of 6 U.S. states)

Algorithm 1 Preliminary Toponym Disambiguation in CBH

Require: Document \mathcal{D} **Require:** Sequence of toponyms T

```
1:  $resolution \leftarrow \emptyset$ 
2: for toponym  $t_i$  in  $T$  do
3:    $l_i \leftarrow \text{NIL}$ 

4:   for interpretation  $l_{i,j}$  in  $L_i$  do
5:      $\mathcal{H}_{i,j} \leftarrow \text{RetrieveHierarchy}(l_{i,j})$ 
6:      $node \leftarrow \text{LookUp}(\text{parent}[l_{i,j}], \mathcal{H}_{i,j})$ 
7:      $score \leftarrow 0$ 
8:     while  $node \neq \text{NIL}$  do
9:       for  $m_h$  in  $\text{Mentions}(node)$  do
10:        for  $m_l$  in  $\text{Mentions}(l_{i,j})$  do
11:           $similarity \leftarrow \max(similarity, \frac{1}{\text{TD}(m_h, m_l)})$ 
12:        end for
13:         $score \leftarrow score + similarity$ 
14:      end for
15:       $node \leftarrow \text{parent}[node]$ 
16:    end while

17:    if  $\text{confidence}[l_i] < score$  then
18:       $l_i \leftarrow (l_{i,j}, score)$ 
19:    else if  $\text{confidence}[l_i] = score$  then
20:      if  $\text{population}[l_i] < \text{population}[l_{i,j}]$  then
21:         $l_i \leftarrow (l_{i,j}, score)$ 
22:      end if
23:    end if
24:  end for
25:   $resolution \leftarrow resolution \cup (t_i, l_i)$ 
26: end for
27: return  $resolution$   $\triangleright$  A preliminary mapping from  $T$  to location interpretations
```

can be huge and enumerating all combinations may not be feasible. To be able to compute probabilities of the hypotheses, we perform a preliminary location disambiguation [197]. This procedure, shown in Algorithm 1, leverages a heuristic to resolve toponyms. Consider a location interpretation $l_{i,j}$ of toponym t_i . The mentions of the ancestors in $l_{i,j}$'s spatial hierarchy (line 5; the hierarchies can be obtained from gazetteer \mathcal{G}) can be used as clues to resolve toponym t_i . The closer an ancestor mention is, the more chance that particular interpretation has to get selected.

For example, toponym *Edmonton* refers to 6 different locations. Provided that it co-occurred with either *Alberta* or *Canada*, we can pinpoint it (*i.e.* the city of *Edmonton* located in *Canada*). For each toponym t_i , the preliminary disambiguation measures a score for each interpretation $l_{i,j}$ (lines 8-16) and picks the interpretation with maximum score (lines 17-18) and in case of tie, the most populous interpretation is selected (lines 19-23). The score is acquired by finding the maximum similarity between $l_{i,j}$ mentions and its ancestors' mentions; similarity here is the inverse of term distance (line 11), as used by Yu and Rafiei [197].

Preliminary disambiguation works poorly in cases where no mentions of locations in spatial hierarchy exist in the document. For instance, suppose we find toponyms *Toronto*, *London*, and *Kingston* in an article. Though, humans can recognize that these cities are presumably located in *Ontario*, *Canada*, preliminary resolution is unable to find any clues for disambiguation and as a result, assigns the toponyms to the interpretation with the highest population (*i.e.* *Toronto* \mapsto *Canada*, *London* \mapsto *England*, and *Kingston* \mapsto *Jamaica*).

The result of the initial phase can be augmented by incorporating context-related features into the resolution process. Our CBH model proceeds to compute probabilities for the two hypotheses. The method operates at each administrative division separately since toponyms may lie in disparate division levels. Hence, the method begins the disambiguation process from the lowest division and furthers the process until all toponyms are resolved.

The geo-centre inheritance indicates that the location interpretation of a toponym can be drawn from the geographic scope of the document. The entities (*i.e.* people, locations, and organizations) used in an article, can ascertain a location interpretation to which the article is geographically relevant [7]. This location defines the geographic scope of the document.

Based on the inheritance hypothesis, the toponyms mentioned in a document are more likely to be part of or under the same administrative division as the geographic

scope of the document. This makes sense due to the *spatial minimality* property. Therefore, we first estimate the geographic scope of the document via a probabilistic model. In particular, for toponym t_i at division d , the probability of $l_{i,j}$ being the correct interpretation is

$$P_{\text{inh}}^{(d)}(l_{i,j}|\mathcal{D}, t_i) = \frac{\text{tf}(\text{anc}_d(l_{i,j}))}{\sum_{p=1}^{n_i} \text{tf}(\text{anc}_d(l_{i,p}))} \quad (4.1)$$

where anc_d returns the ancestor of an interpretation at division d and $\text{tf}(w)$ computes the term frequency in the document. Each location interpretation here is extended to include its corresponding spatial hierarchy. For example, interpretations of toponym *Paris* are represented as

$$\{ [Paris \rightsquigarrow Ile-de-France \rightsquigarrow France], \\ [Paris \rightsquigarrow LamarCounty \rightsquigarrow Texas \rightsquigarrow US], \dots \}$$

The second hypothesis, namely near-location hypothesis, relies upon the toponyms mentioned in the vicinity of a toponym. Toponyms nearby a toponym can be linked to one another primarily because of *object/container* and *comma group* relationships they possibly have [110]. According to this hypothesis, the closer toponym s to toponym t , the stronger evidence toponym s is to disambiguate toponym t . This is why, in this hypothesis, we compute the term distance between toponyms as a measure of similarity to estimate probabilities. In effect, for toponym t_i at division d , the probability of $l_{i,j}$ being the correct interpretation is

$$P_{\text{near}}^{(d)}(l_{i,j}|\mathcal{D}, t_i) = \frac{\text{sim}(t_i, \text{anc}_d(l_{i,j}))}{\sum_{p=1}^{n_i} \text{sim}(t_i, \text{anc}_d(l_{i,p}))} \quad (4.2)$$

where $\text{sim}(v_1, v_2)$ is the similarity function between terms v_1 and v_2 as demonstrated below:

$$\text{sim}(v_1, v_2) = \frac{1}{\min_{w_i \in M(v_i)} \{\text{TD}(w_1, w_2)\}} \quad (4.3)$$

where $\text{TD}(w_1, w_2)$ is the distance between indices of w_1 and w_2 and $M(v)$ is a set containing the mentions of term v in document \mathcal{D} .

Now, we combine $P_{\text{inh}}^{(d)}$ and $P_{\text{near}}^{(d)}$ to incorporate both premises into the model. The final context-bound model is regarded as a weighted linear function of the two probabilities:

$$P_{\text{CB}}^{(d)}(l_{i,j}|\mathcal{D}, t_i) = J^{(d)}(\mathcal{D}, t_i) \cdot P_{\text{near}}^{(d)}(l_{i,j}|\mathcal{D}, t_i) + (1 - J^{(d)}(\mathcal{D}, t_i)) \cdot P_{\text{inh}}^{(d)}(l_{i,j}|\mathcal{D}, t_i) \quad (4.4)$$

The coefficient $J^{(d)}(\mathcal{D}, t_i)$ is obtained via Shannon Entropy of the vector induced by near-location probabilities for toponym t_i with respect to $l_{i,j}$ for all values of j .

The resolution is undertaken through maximum likelihood estimation over the probability in Equation (4.4). The final computed probability can be considered as confidence score.

Algorithm 2 CBH Resolution

Require: Document \mathcal{D} and sequence of toponyms T

Require: A mapping from T to location interpretations

```

1: resolution ← PreliminaryResol(D,T)                                ▷ Algorithm 1
2: for  $k=1$  to maxIterations do
3:   for division  $d$  in {County, State, Country} do
4:     for toponym  $t_i$  in  $T$  do
5:        $\ell_i \leftarrow \operatorname{argmax}_j \{P_{\text{CB}}^{(d)}(l_{i,j}|\mathcal{D}, t_i)\}$            ▷ Refer to Eq. (4.4)
6:       resolution ← resolution  $\cup (t_i, \ell_i)$ 
7:     end for
8:   end for
9: end for

```

In summary, the CBH resolution method is illustrated in Algorithm 2. The approach starts with a preliminary resolution, followed by a hypotheses assessment to rectify results from the initial resolution. The hypotheses model computes the probabilities for each division separately to ensure the model can afford toponyms in all levels of dispersion. Once the modification process finished, the algorithm repeats for another iteration since altering the resolution of a toponym may affect other disambiguated toponyms. Our experiments show that CBH often takes two iterations to complete. However, in some cases, the modification step never terminates. Specifically, consider the following sentence, an excerpt from a news article:

“... London’s Heathrow, one of the world’s busiest travel hubs.”⁸

London and *Heathrow* are recognized as toponyms. Because no notion of ancestors in the spatial hierarchy can be found, the initial resolution favors the highest population interpretation (i.e., *London* \mapsto *England* and *Heathrow* \mapsto *Florida, US*). In the next step, the hypotheses model maps *London* to a place in United States because the other toponym is located in United States. Accordingly, *Heathrow* is assigned to the airport in England. After the first iteration, the resolution is changed to $\{London \mapsto US, Heathrow \mapsto England\}$. Conversely, the second iteration would alter the results to $\{London \mapsto England, Heathrow \mapsto US\}$; the algorithm is now trapped in an infinite loop. This is why, we introduce *maxIterations* parameter to eschew these circumstances. While CBH fails to successfully resolve toponyms in such cases, the approach, described in the next section, can address this shortcoming.

4.3.2 Spatial-Hierarchy Sets

The spatial minimality property (noted by Leidner [102]) leads us to another resolution method called *Spatial-Hierarchy Sets (SHS)*. This method is grounded on containment and sibling relationships that are likely to exist among toponyms in a document. Consider a non-disjoint partitioning of the universe of locations (in a gazetteer) where locations with similar or related interpretations (e.g. those under the same administrative division or within a close proximity) form a partition. Since toponyms in a document tend to refer to geographically related locations, and those locations are more likely to be in the same partitions than different partitions, we want to find a small set of partitions that cover all toponyms; this can be modeled as a conflict-free covering problem. Conflict-free covering refers to the traditional set cover problem where each element must be covered by at most one set in the answer. The covering needs to be conflict-free due to *one-sense-per-referent* property. We formally define conflict-free covering as an instance of the conflict-free coloring of

⁸<http://money.cnn.com/2016/12/14/news/companies/british-airways-ba-strike-christmas>

regions [71].

Conflict-free Covering Problem

Given a finite family of finite sets \mathcal{S} where each set S_i is associated with a non-negative weight w_i and a universal set \mathcal{U} containing all the elements from the sets, we seek to find a collection of sets, namely \mathcal{A} , with minimum weight such that their union becomes \mathcal{U} while each element is covered by at most one set in \mathcal{A} .

We formulate toponym resolution by conflict-free covering problem as the following:

1. Each parent with all its children form a set of related interpretations. Let \mathcal{S} denote the collection of all such sets that can be constructed. Each parent appears in a set with its children, hence the size of \mathcal{S} is the same as the number of parents with non-zero children. Algorithm 3 depicts the details of generating \mathcal{S} .
2. Recall that T denotes the set of toponyms in document \mathcal{D} as defined in Section 4.2. We say a set in \mathcal{S} covers a toponym in T , if the set contains the surface text of the toponym. We want to select sets in \mathcal{S} that cover all toponyms in T . Our goal is to minimize the number of interpretations (spatial minimality) by selecting as few sets in \mathcal{S} as possible.
3. Let us form a color class for each toponym. The color class for a toponym includes all possible interpretations of the toponym. For example, *Texas* is a color class which includes all places that can resolve *Texas*. We want to avoid selecting multiple interpretations for the same toponym. That means, we do the selection in (2) with the constraint that no more than one color or interpretation can be selected for each toponym.

In the special case where the color classes are empty (*i.e.* no constraint on colors), the problem becomes the classic set cover, which is NP-complete. This means that

Algorithm 3 Spatial-Hierarchy Set Generation

Require: Document \mathcal{D} and sequence of toponyms T

```
1:  $\mathcal{S} \leftarrow \emptyset$ 
2:  $P \leftarrow \emptyset$ 

3: for toponym  $t_i$  in  $T$  do
4:   if  $name[t_i]$  in  $P$  then
5:     skip  $t_i$ 
6:   end if
7:   for interpretation  $l_{i,j}$  in  $L_i$  do
8:     if  $parent[l_{i,j}]$  in  $\mathcal{S}$  then ▷ Checks whether the set exists
9:        $AddChild((l_{i,j}, \mathbf{true}), \mathcal{S}[parent[l_{i,j}]])$ 
10:    else
11:      ▷ The new set is a tree whose root is  $parent[l_{i,j}]$ 
12:      ▷ The boolean values represent mentioned flags
13:       $\mathcal{S} \leftarrow \mathcal{S} \cup \{(parent[l_{i,j}], \mathbf{false}) \rightarrow (l_{i,j}, \mathbf{true})\}$ 
14:      if  $l_{i,j}$  in  $\mathcal{S}$  then
15:         $mentioned[\mathcal{S}[l_{i,j}]] \leftarrow \mathbf{true}$ 
16:      end if
17:    end if
18:     $P \leftarrow P \cup name[l_{i,j}]$ 
19:  end for
20: end for
21: return  $\mathcal{S}$  ▷ A collection of spatial hierarchy sets
```

existing methods approximate the optimal solution. We leverage a greedy approach [172] to solve the problem. Although the greedy approach gives an approximate answer to the problem in general, our experiments reveal that such answer yield a competitive performance.

However, this model suffers from some deficiencies, even if an optimal solution is reached. A problem with this formulation is that we cannot have *Montreal, Quebec* and *Windsor, Ontario* in the same text (or they will not be resolved correctly) because *Windsor* is also a town in *Quebec*. These are cases where the hypotheses model, namely CBH, can better resolve. Furthermore, there may be circumstances that similar toponyms may appear in more than one sets and yet, we cannot favor one set to another. Suppose we have a document where only *Georgia* and *Turkey* are mentioned. Two sets, $\{Georgia(\text{city}) \rightsquigarrow Texas(\text{state}), Turkey(\text{city}) \rightsquigarrow Texas(\text{state})\}$

and $\{Georgia(\text{country}) \rightsquigarrow World, Turkey(\text{country}) \rightsquigarrow World\}$, would emerge in \mathcal{S} . Without any additional information, such as document source, even humans cannot choose the correct interpretation. SHS selects the most populated set as a rule of thumb in these cases.

4.3.3 Context-Hierarchy Fusion

While the Spatial-Hierarchy Sets approach guarantees the minimality properties, it fails to select between identical structures (*e.g.* the *Georgia* and *Turkey* case) mostly because it does not delve into other context-related features of the document. On the other hand, the Context-Bound Hypotheses model benefits from term frequency and term distance features of the context. Notwithstanding the situations like *Georgia* and *Turkey*, using other context sensitive information alleviates the disambiguation process in most cases. For example, toponyms *London*, *Aberdeen* and *Edinburgh* have interpretations located in *Canada* and SHS resolves them to the corresponding interpretations in *Canada* to preserve minimality. Even the presence of toponym *England* does not change the result because *Aberdeen* and *Edinburgh* located in *Scotland* and we still need to pick two sets to attain the correct resolution.

Consequently, merging SHS and CBH method allows us to take advantage of both methods at the same time. *Context-Hierarchy Fusion (CHF)* method chooses an interpretation from CBH only if the confidence score is higher than a threshold τ . Otherwise, it resolves toponyms using SHS.

4.4 Experiments

In this section, we conduct extensive experiments to evaluate our methods⁹ and to assess their performance under different settings. The particular questions to be investigated are:

⁹The source code and the annotated dataset is available at <https://github.com/ehsk/CHF-TopoResolver>

1. Given that CBH comprises different steps and components, how much does inheritance and near location hypothesis improve upon the preliminary location disambiguation?
2. How sensitive is Context-Hierarchy Fusion to the value of the threshold and if there are some sweet spots?
3. How accurate is the proposed method, compared to the state-of-the-art supervised and unsupervised methods as well as commercial systems?
4. How does the proposed method compare to the state-of-the-art supervised method in terms of the generality of the model on unseen data?
5. When is an unsupervised technique expected to surpass supervised methods?

For (3), we compare the performance of our method to that of the state-of-the-art methods as well as commercial systems: Yahoo! YQL Placemaker, OpenCalais and Google Cloud Natural Language API. The details of these proprietary products have not been made public. However, these systems can be accessed through public Web APIs at a relatively liberal rate limit, which enable us to automatically test their geotagging process on our datasets.

In our evaluation setting, we apply two methods for toponym recognition. First, we assume that the recognition phase is flawless, which is displayed as *Resol*. In this method, the annotated toponyms without latitude/longitude are fed to the underlying resolution method. These experiments are conducted to compare our methods to resolution methods such as TopoCluster [41]. Second, we employ Stanford NER [55] to tag locations, which is shown by *GeoTag*. We run *GeoTag* experiments to draw a comparison with systems performing both recognition and resolution including closed-source products and Adaptive [109].

4.4.1 Datasets

In order to evaluate our toponym resolution methods, gold data corpora are required, in which all occurrences of geographic names and phrases have been manually annotated. In our experiments, we exploit three annotated datasets:

- **TR-News:** We collected this dataset from various global and local News sources. We obtained news articles from several local news sources to include less dominant interpretations of ambiguous locations such as *Edmonton, England* and *Edmonton, Australia* rather than *Edmonton, Canada* or *Paris, Texas, US* in lieu of *Paris, France*. Additionally, a number of articles from global news sources such as BBC and Reuters have been selected to preserve the generality of the corpus. We manually annotated toponyms in the articles with the corresponding entries from GeoNames. The gold dataset consists of 118 articles.
- **Local-Global Lexicon (LGL):** This corpus was curated by Lieberman *et al.* [107]. It is collected from local news sources and mainly focuses on including ambiguous toponyms and this is why, it is suitable to test toponym resolution systems against geographically localized documents. The dataset is composed of 588 articles from 85 news sources.
- **CLUST:** Lieberman and Samet [108] compiled this dataset from a variety of global and local news sources. CLUST is a large dataset containing 1082 annotated articles.

Table 4.1 summarizes and compares the statistics of these datasets. The median number of toponyms per document in all datasets are close to each other, meaning that the corpora do not differ significantly with one another in terms of the number of toponyms per article.

In addition, the three datasets contain toponyms (roughly 3%) that cannot be found in gazetteer \mathcal{G} , while annotated and linked to an entry in the gazetteer. We

Table 4.1: Corpora used in our experiments

	TR-News	LGL	CLUST
News sources	36	85	352
Documents	118	588	13327
Annotated docs	118	588	1082
Annotated topos	1318	5088	11962
Topos with GeonameID	1274	4462	11567
Distinct topos	353	1087	2323
Median topos per doc	9	6	8
Topos not found in GeoNames	2.7%	3.2%	3.3%
Wikipedia-linked topos	94.3%	94.1%	94.2%

observe that such toponyms fall into one of the following categories: uncommon abbreviations such as *Alta.* stands for *Alberta, Canada*, multi-word places such as *Montreal-Pierre Elliott Trudeau International Airport*, and transliterated place names (*e.g.* city of *Abbasiyeh, Egypt* written as *Abbassiya*).

The test corpora is also analyzed by the location type of their annotated toponyms, as done by Lieberman and Samet [109]. We compute the percentage of each location type for each dataset. As show in Figure 4.1, *LGL* dataset largely consists of small cities, which makes it a challenging test dataset since well-known locations are presumably to be resolved with high precision due to their frequent use in articles. In contrast, *TR-News* and *CLUST* datasets are roughly similar and include countries more than any location type. This denotes that the articles appeared in *TR-News* and *CLUST* are extracted from sources that are aimed at a global audience. These sources usually provide more details for location mentions such as saying *Paris, US* instead of *Paris*. On the other hand, in *LGL*, because the articles are meant to be of use for local audience, the news publishers typically do not state additional information in this regard. Thus, geotagging approaches can be tested against these test

corpora since they span a variety of news sources both globally and locally.

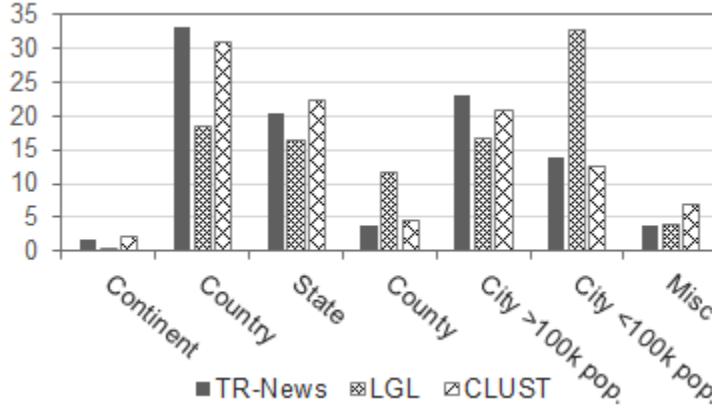


Figure 4.1: Comparative analysis of the test datasets based on location type

4.4.2 Evaluation Metrics

Performance measures in our experiments are *Precision*, *Recall*, F_1 -measure, and mean error distance (M). However, to ascertain whether an interpretation is correctly predicted, we also investigate the error distance between the predicted coordinates and the actual coordinates, as used in numerous studies [29, 41, 173, 102, 159, 109, 8]. This distance enables us to fare various systems against each other since they may select latitude/longitude of locations from different gazetteers or knowledge bases¹⁰. We set the error distance to 10 miles, same as Adaptive method [109], whereas most researches tend to adopt a relaxed threshold (*i.e.* 161 kilometers) [173, 41, 159, 8].

For TopoCluster [41] and the commercial products, on the other hand, we employ a different criteria primarily because the error distance may not be accurate for large areas (even with higher error distance thresholds). Hence, in order to consider whether an interpretation is correctly projected to a coordinate, we check if the predicted interpretation resides in the bounding-box area of the ground truth; here the bounding-boxes of locations are extracted from GeoNames. We did not use this bounding-box grounded accuracy for other methods since most of them rely on the

¹⁰Locations are represented with a single centroid and gazetteers may vary in picking the centroids.

same gazetteer adopted in this work. Although using bound-boxes works in favor of TopoCluster and the proprietary products, the mean error distance fails to precisely mirror the accuracy for these methods since for a prediction deemed as correct based on bounding-boxes, the error distance can still be high.

Furthermore, in *Resol* experiments, we only calculate *Precision* because given a toponym, a resolution method is more likely to map it to an interpretation unless it does not exist in the gazetteer; thus, *Recall* would be approximately analogous to *Precision*. It is also worth mentioning that the mean error distance is only reported in *Resol* experiments and not in *GeoTag* experiments, because the mean error distance cannot be measured for toponyms that are either not identified or falsely detected.

4.4.3 Analysis of Context-Bound Hypotheses

As discussed in Section 4.3.1, Context-Bound Hypotheses commences with a preliminary toponym disambiguation, followed by estimating two probabilities for inheritance and near-location hypotheses. In this section, we evaluate the preliminary phase and see whether the modification phase by Context-Bound hypotheses alleviates the resolution performance. Moreover, we study the role of the hypotheses in CBH by removing one of them at a time and measuring the performance. This experiment is conducted on the *TR-News* dataset in both *Resol* and *GeoTag* modes.

As shown in Table 4.2, taking both hypotheses into account complements the preliminary disambiguation, though the improvement does not seem considerable (slightly higher than 1% in F_1 -measure) in both *GeoTag* and *Resol* experiments.

Additionally, the near-location hypothesis contributes to the improvement more than the inheritance hypothesis. This is largely because the inheritance hypothesis estimates probabilities using term frequency. In cases where two locations are mentioned as frequent as each other, term frequency does not seem accurate. For example, consider the toponym *Edmonton*, which can be located in either *Canada* or *Australia* in a document where *Australia* and *Canada* appear twice each. This

Table 4.2: Detailed analysis of Context-Bound Hypotheses (CBH) on *TR-News* dataset

	P_{Resol}	P_{GeoTag}	R_{GeoTag}	$F_{1-\text{GeoTag}}$
Preliminary	78.0	73.4	52.1	60.9
Inheritance	78.1	73.6	52.0	61.0
Near-location	79.0	73.9	52.3	61.2
CBH	79.2	74.9	53.0	62.1

results in the same score for both interpretations and a decision would be made by population size. Term distance, however, can help better in this case, denoting that the closer mention is more likely to be the correct interpretation. Nonetheless, we still need both hypotheses since the results are improved by putting near-location and inheritance together.

4.4.4 Fusion Threshold Study

In Context-Hierarchy Fusion, explained in Section 4.3.3, choosing an appropriate value for the threshold can be crucial in the resolution the performance. In this experiment, we vary the threshold τ to study its effect on performance. According to the results shown in Figure 4.2, we can identify a sweet spot when CHF achieves the best performance on all three datasets; this happens when τ falls between 0.5 and 0.6; we set τ to 0.55 in our experiments.

Also, we can see a mild spike in F_1 -measure at $\tau = 1$ in the *LGL* curve, which can be attributed to the localized content of the dataset. In particular, SHS (at $\tau = 1$, CHF is analogous to SHS) works better on *LGL* since locations in *LGL* are not mentioned frequently alongside their corresponding spatial hierarchy ancestors. As discussed in Section 4.3.1, CBH needs to spot the mentions of these ancestors in documents (containment relationship) in order to generate a more accurate resolution, whereas SHS does not rely solely on containment relationships. It also takes sibling

relationships into account, and as a result, merging SHS and CBH does not seem to be effective on *LGL*.

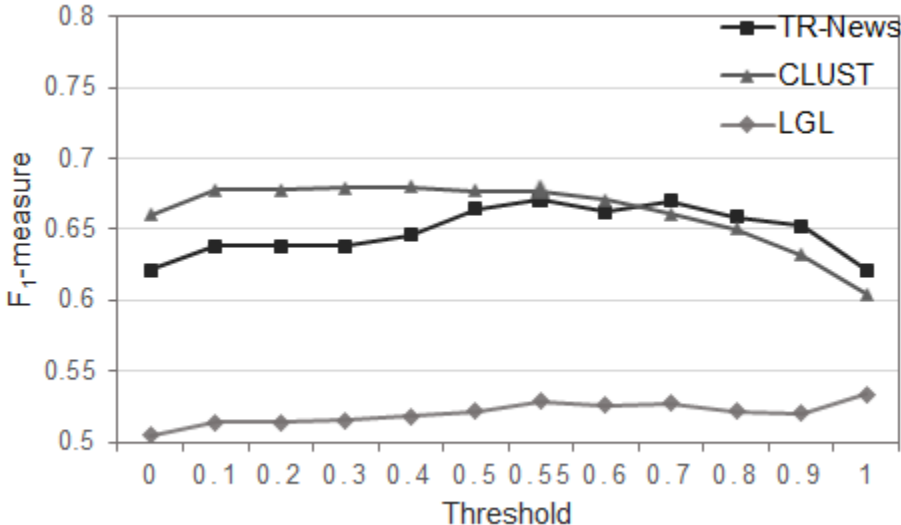


Figure 4.2: F_1 -measure vs. threshold τ for Context-Hierarchy Fusion method on *TR-News* dataset. At $\tau = 0.55$, CHF achieves the best F_1 -measure on all three corpora.

4.4.5 Resolution Accuracy

In this section, we measure the performance of our proposed methods and compare them with other resolution techniques. The methods presented in this work are Context-Bound Hypotheses (CBH), Spatial-Hierarchy Sets (SHS) and Context-Hierarchy Fusion (CHF). We compare the results with two prominent systems: TopoCluster [41], the state-of-the-art unsupervised model, and Adaptive [109], the state-of-the-art supervised model. The source code of TopoCluster was available online, so we were able to test the method on our datasets. However, in order to test the Adaptive classifier, we implemented the supervised method¹¹, albeit without two features, namely *dateline* and *locallex*; this was because for *locallex*, the authors used an annotated dataset containing the expected audience location of the news sources and also, *dateline* required a general location for each article which was not available for

¹¹Since we did not have access to the source code.

most articles in the test corpora. The modified version is named *CustomAdaptive* in our results. We follow the same parameter setting of the original Adaptive [109] and perform 10-fold cross validation to test CustomAdaptive.

Table 4.3 and Table 4.4 illustrate the evaluation results. CHF produces the best performance among our proposed methods on *CLUST* and *TR-News* and SHS beats the other proposed techniques on *LGL*. Among all listed methods, CustomAdaptive shows the highest performance. We also report recall, to make a comparison with the original Adaptive method [109].

While commercial products produce high precision, their recall is lower than our proposed methods in all cases except for Yahoo! YQL Placemaker. Placemaker yields the best results among the commercial products and achieves higher overall performance than our methods. On the other hand, OpenCalais is able to recognize toponyms as locative expressions. For instance, it identifies *the Kenyan captial* rather than just *Kenyan*. However, we observe that sometimes it fails to detect a full location phrase; for example, only *Toronto* in *Greater Toronto Area* is detected¹². Further, Google Cloud Natural API offers an entity extraction service, which focuses highly on recognition of named entities¹³. The system links extracted entities to their corresponding Wikipedia articles and provides no additional information about geographic coordinates of location entities. Therefore, the geographical information of locations can only be derived from Wikipedia for this product. According to Table 4.1, nearly 94% of toponyms in each dataset have Wikipedia articles¹⁴, but not all Wikipedia articles contain spatial coordinates of locations, which is partly attributed to a poor recall in our experiments. Thus, we can see why entity linking approaches cannot be exploited for toponym resolution.

We run *Resol* experiments to analyze TopoCluster [41] since it is a resolution

¹²We count these as correct resolutions unless they fall outside the bounding box of the annotated toponym.

¹³Google Cloud Natural API extracts locative expressions in any form in addition to proper names like *family home* and *suburb*.

¹⁴GeoNames keeps record of Wikipedia URLs for each location.

method. DeLozier *et al.* stipulated that TopoCluster performs best when integrated with a gazetteer; this is why, the integrated version, called *TopoClusterGaz*, is adopted throughout this experiment. The results are presented in Table 4.3 and Table 4.4 (P_{Resol} and M_{Resol} columns). According to our results, CHF outperforms TopoCluster on all three datasets. Moreover, DeLozier *et al.* [41] set the error distance threshold for TopoCluster to 161 kilometers and achieved an accuracy of 71.4% on *LGL*¹⁵, whereas under the same setting, CHF reaches 71.2% on *LGL*, which is marginally lower than TopoCluster.

Besides accuracy, the mean error distance is also measured in our *Resol* experiments¹⁶. Among the unsupervised methods, CBH stands out with the lowest error. CHF is close to CBH with its error not exceeding 40km. This difference stems from SHS impacting CHF because when a toponym is projected to an incorrect location by SHS, the mapped location is more likely located in a country different than the ground truth.

4.4.6 Unseen Data Analysis

Supervised techniques benefit from the knowledge gained in the training phase and if there is an overlap between the training data and the test data, then the prediction can be counted as overly optimistic. Domingos [45] emphasizes that generalization is achieved by a separation of the training data and the test data [45]. This is why, we study the effect of the overlap between training and test datasets on F_1 -measure. For this purpose, CustomAdaptive classifier was trained on *CLUST* dataset (the trend does not vary significantly if the classifier trained on *LGL*) and tested against *TR-News*. We define the overlap ratio measure as the number of toponyms per article in test data, which has also been appeared in the training data. We can channel overlap ratio through trimming off articles from test data and measure performance on the

¹⁵Among the datasets used in TopoCluster paper [41], *LGL* is the only dataset to which we have access

¹⁶Mean error distance for TopoCluster in *LGL* is derived from the original paper [41].

Table 4.3: Performance results in *GeoTag* and *Resol* experiments on LGL and CLUST. The best results in each category are bolded.

Method	LGL					CLUST				
	P	R	F_1	P_{Resol}	M_{Resol}	P	R	F_1	P_{Resol}	M_{Resol}
<i>Unsupervised</i>										
CBH	66.8	40.6	50.5	68.6	760	80.6	55.8	66.0	81.5	709
SHS	69.7	43.3	53.4	68.3	1372	72.8	51.6	60.4	71.1	1521
CHF	68.5	43.1	52.9	68.9	818	80.6	58.4	67.7	81.0	788
TopoCluster [41]	-	-	-	59.7	1228	-	-	-	77.1	769
<i>Supervised</i>										
Adaptive [109]	-	58.7	-	94.2	-	-	61.8	-	96.0	
CustomAdaptive	79.2	48.5	60.2	88.3	679	89.8	57.9	70.4	93.4	504
<i>Commercial</i>										
Placemaker	73.5	48.6	58.5	-	-	87.4	61.1	71.9	-	-
OpenCalais	77.1	28.9	42.1	-	-	87.5	48.5	62.4	-	-
GoogleNL+Wiki	80.5	34.0	47.8	-	-	82.8	39.2	53.2	-	-

Table 4.4: Performance results in *GeoTag* and *Resol* experiments on TR-News. The best results in each category are bolded.

Method	P	R	F_1	P_{Resol}	M_{Resol}
<i>Unsupervised</i>					
CBH	74.9	53.0	62.1	79.2	869
SHS	73.8	53.6	62.1	69.9	2305
CHF	79.3	58.2	67.1	80.5	942
TopoCluster [41]	-	-	-	68.8	1422
<i>Supervised</i>					
CustomAdaptive	83.8	74.9	79.1	90.5	573
<i>Commercial</i>					
Placemaker	80.8	63.0	70.8	-	-
OpenCalais	81.3	48.5	61.2	-	-
GoogleNL+Wiki	80.2	38.4	51.9	-	-

trimmed test data. Figure 4.3 plots F_1 -measure against the overlap ratio. The unsupervised method surpasses the supervised method when the overlap ratio is less than 60% (when the overlap ratio is at 0.6, CHF still outperforms CustomAdaptive with a 1% margin). This observation confirms that the unsupervised technique, namely CHF, can handle unknown data better than the supervised method, namely Adaptive (CustomAdaptive implementation).

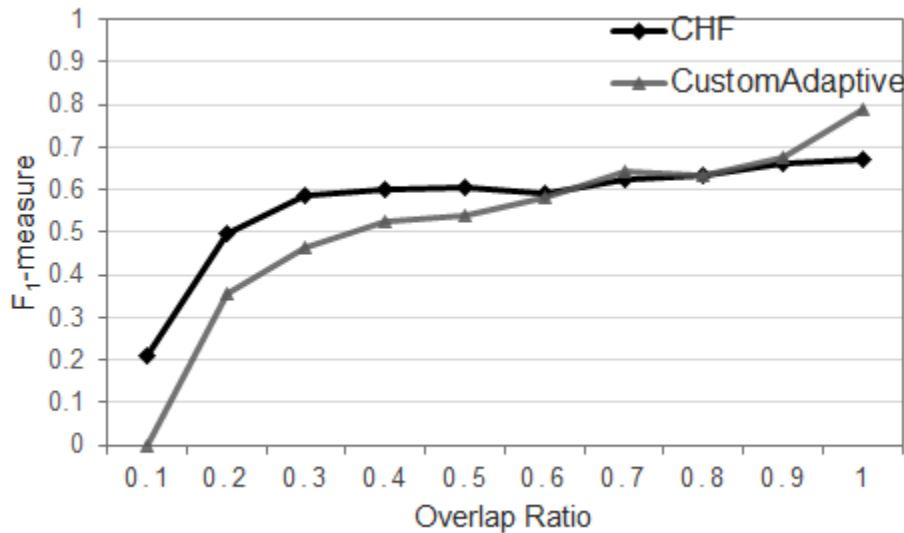


Figure 4.3: F_1 -measure of CustomAdaptive trained on CLUST and CHF when overlap ratio varies. CHF yields a better performance than CustomAdaptive when overlap between training data and test data is lower than 60%.

4.5 Related Works

Numerous studies have been conducted and much progress has been made on the task of disambiguating location mentions. The existing approaches in the literature may be grouped into (1) unsupervised and rule-based, (2) supervised, and (3) those based on some knowledge bases. However, a plethora of methods leverage a mixture of techniques. For example, DeLozier *et al.* [41] proposed an unsupervised toponym resolution method that leverages geographical kernels and spatially annotated Wikipedia articles. Also, Lieberman and Samet [109] presented a supervised technique that uses

both geographical distance and additional knowledge like gazetteers and document source to disambiguate toponyms.

Unsupervised and rule-based methods: In unsupervised resolution, various techniques have been studied. Map-based methods create a representation of all referents on a world map and apply techniques such as geographical centroid detection and outlier elimination to estimate the target of a toponym [102]. Moncla *et al.* [134] introduced a map-based technique where density-based clustering was carried out to detect outliers. Buscaldi [24] found that map-based techniques face difficulties in grounding toponyms in a document when they are spatially far from each other. Rule-based and heuristic-based methods also have been adopted in the literature [102, 6]. For instance, the presence of “Canada” in text *London, Canada* may help disambiguate *London*. However, finding a set of rules to cover all cases in natural language text seems to be arduous.

Approaches using knowledge bases: Wikipedia has been integrated as a knowledge base into more recent toponym disambiguation techniques [8, 41, 174, 173, 159]. Ardanuy and Sporleder [8] tackled toponym disambiguation in multilingual retrospective articles. They built a model to distill semantic features from Wikipedia information such as page title and article body. Speriosu and Baldrige [173] found that non-spatial words impart useful information to disambiguate toponyms and they propose likelihood models that are obtained from Wikipedia. DeLozier *et al.* [41] proposed TopoCluster, which does not rely on gazetteers to resolve toponyms, to address cases where location mentions are not found in gazetteers. They constructed a geographical language model to capture geographical senses of words using Wikipedia pages of locations. However, adding gazetteer information to TopoCluster, namely *TopoClusterGaz*, yields a better performance. Less known toponyms are not expected to be found in Wikipedia; they can introduce challenges and hinder the performance

of this method.

Supervised methods: Many classification techniques have been proposed for geo-tagging purposes including Bayesian [1], random forests [109], RIPPER rule learner [60] and SVM [60, 128]. The features extracted for these classifiers can be grouped into context-free and context-sensitive features [109]. Context-free features typically include heuristics and information from external sources such as knowledge bases and gazetteers and may include, for example, population [109] and location type [60]. Context-sensitive features are obtained from documents where toponyms are mentioned. Melo and Martins [128] used normalized TF-IDF document vectors over curvilinear and quadrilateral regions on Earth’s surface. The adaptive method, proposed by Lieberman and Samet [109], casts geographical proximity and sibling relationship among interpretations in a context window as features. GeoWhiz [1] aggregates several likelihoods based on observations in training data. For instance, largely populated places are more likely estimated as their prominent interpretation. The suggested method by Santos *et al.* [159] consolidates information from Wikipeage pages of locations to compute several similarity and geographical features (context-free features) and performs a nearest neighbor search using locality-sensitive hashing (LSH) to resolve locations.

Other more general related work: Entity disambiguation (also known as entity linking) [106, 115, 165, 75, 58] is related to toponym resolution. Linking named entities (*i.e.*, people, organizations, and locations) to their corresponding real world entities in a knowledge base subsumes toponym disambiguation. Nonetheless, geographical features of location entities are neglected by these systems [173] and thus, geographically specialized methods for resolving toponyms are still needed to map locations to their corresponding geographic footprint.

Another line of research pertinent to this work is location disambiguation in social

media. The related work in this area may incorporate user profile data and social network information as well as natural language processing tools and gazetteers to tackle this task [79, 136]. Flatow *et al.* [56] proposed a method that learns geo-referenced n -grams from training data to perform geotagging on social messages. Use of words that are endogenous to social media are considered as an inherent hurdle here. Moreover, social media content have deficient orthographic structure and lack context, which bring even more complexities to toponym resolution in social media [51, 152, 125].

4.6 Conclusions

We studied toponym resolution and proposed two novel unsupervised models and a mixture model, namely CHF, to address the problem. We investigated the effectiveness of the proposed methods with other techniques. Our evaluations show that the Context-Hierarchy Fusion method outperforms TopoCluster, the state-of-the-art unsupervised method, in terms of precision. The performance of supervised techniques exceeds that of our proposed methods (as expected), nonetheless, we have shown that the state-of-the-art supervised classifier, called Adaptive, highly relies on the training data and Context-Hierarchy Fusion can handle unseen toponyms better.

The future work may investigate other mixture models and a better understanding of when one or both of supervised and unsupervised methods are expected to perform not so well. Moreover, the correlations among the bounding-boxes of toponyms in an article can be studied to augment the resolution, considering the gazetteer are endowed with bounding-box of locations for this purpose [171]. Another direction is understanding the differences between short and long text as far as toponym resolution is concerned and the challenges each pose.

Chapter 5

Document-level Reasoning: A Hidden Challenge in Open-domain QA Benchmarks

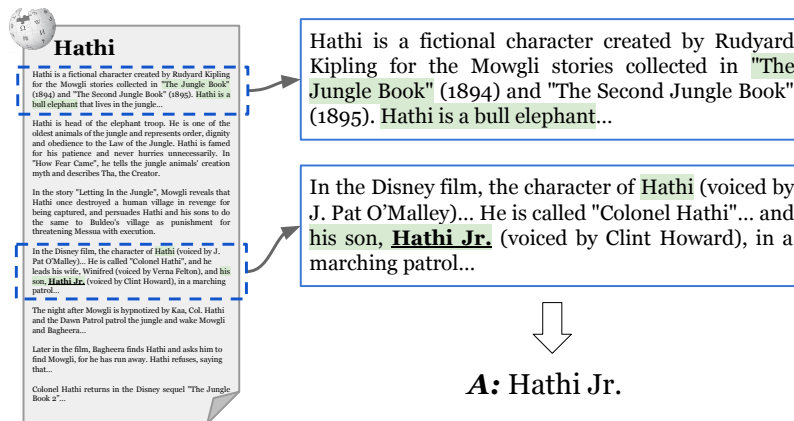
The research community, especially in IR and NLP, rapidly advances by developing models that achieve remarkable performance on established benchmarks. Nonetheless, benchmarks per se are often taken for granted. Unfortunately, studies that aim at critical analysis of benchmarks are far fewer than the number of existing benchmarks. Luckily, some studies published in recent years thoroughly inspect well-known datasets in computer vision [64, 193, 89, 28] and in NLP [16, 126, 176, 140]. In this chapter, I pursue a similar goal on open-domain QA benchmarks towards reliable evaluation (D4), a key requirement specified in Section 1.2. More precisely, my main focus is to highlight an underlying pitfall in existing open-domain QA benchmarks, as the saying goes “you can’t improve what you don’t measure.”¹

5.1 Introduction

Answering information-seeking questions over a massive collection of documents, known as *Open-domain Question Answering (QA)*, has been a long-standing goal of NLP research. Retrieving candidate documents that contain potential answer(s) is at the heart of open-domain QA. A wide range of retrieval models have been adopted

¹Often attributed to Peter Drucker, an influential figure in modern management studies.

Q: What is the baby elephant's name in Jungle Book?



Hathi

Hathi is a fictional character created by Rudyard Kipling for the Mowgli stories collected in "The Jungle Book" (1894) and "The Second Jungle Book" (1895). Hathi is a bull elephant that lives in the jungle...

Hathi is head of the elephant troop. He is one of the oldest animals of the jungle and represents order, dignity and obedience to the Law of the jungle. Hathi is famed for his patience and never hurries unnecessarily. In "How Four Came", he tells the jungle animals' creation myth and describes The, the Creator.

In the story "Letting in the Jungle", Mowgli reveals that Hathi once destroyed a human village in revenge for being captured, and persuades Hathi and his sons to do the same to Badno's village as punishment for threatening Messua with execution.

In the Disney film, the character of Hathi (voiced by J. Pat O'Malley)... He is called "Colonel Hathi", and he leads his wife, Winifred (voiced by Yvonne Peck), and his son, **Hathi Jr.** (voiced by Clint Howard), in a marching patrol...

The night after Mowgli is hypnotized by Kaa, Col. Hathi and the Dawn Patrol patrol the jungle and wake Mowgli and Bagheera...

Later in the film, Bagheera finds Hathi and asks him to find Mowgli, for he has run away. Hathi refuses, saying that...

Colonel Hathi returns in the Disney sequel "The Jungle Book 2"...

Hathi is a fictional character created by Rudyard Kipling for the Mowgli stories collected in "The Jungle Book" (1894) and "The Second Jungle Book" (1895). Hathi is a bull elephant...

In the Disney film, the character of Hathi (voiced by J. Pat O'Malley)... He is called "Colonel Hathi"... and his son, **Hathi Jr.** (voiced by Clint Howard), in a marching patrol...

↓

A: Hathi Jr.

Figure 5.1: An example question, taken from the Natural Questions-OPEN test set [101], that highlights the importance of document-level reasoning in retrieving passages.

for this purpose, from sparse retrievers such as BM25 [154] to dense retrievers [90] and retrieval-augmented models [103, 70]. Yet, retrieval is invariably conducted over a collection of passages. In particular, the standard practice is to split documents (*e.g.* Wikipedia articles, news articles) into fixed-length passages [185]. The ubiquity of passage retrieval is mainly due to the limited context size of deep learning models [43, 3]. Moreover, existing models often struggle in capturing long-range dependencies within documents [92]. In open-domain QA, it is well documented that passage retrieval is effective on several benchmarks [35, 185, 90].

Notwithstanding the success of passage retrieval in open-domain QA, such passage-level treatment of documents largely ignores document-level evidence that may be essential in finding candidate answers. Documents are written in a logically-structured manner and follow a cohesive narrative [83]. By carving their discourse into passages, the underlying relationship among different parts of documents (*e.g.* coreferences) is no longer upheld. These issues introduce additional challenges that can impede models from answering some questions. For example, in Figure 5.1, the key information to answer the question is dispersed in two paragraphs that are distant from each another.

In this work, we inspect open-domain QA datasets to identify questions that require document-level reasoning to answer. We conduct document-level retrieval on three widely adopted open-domain QA datasets: Natural Questions-OPEN [101], TriviaQA [85], and WebQuestions [19]. We then find questions for which document-level retrieval outperforms passage-level retrieval. Our observation, consistent with the literature, is that passage retrieval provides better overall performance than document retrieval on these datasets. However, on 325 questions, passage retrieval either completely fails or underperforms document retrieval. We manually audit these questions to determine whether document-level information is actually required for them. Our analysis reveals that these benchmarks are heavily skewed toward questions where passage-level information is sufficient. Despite the prevalence of such questions, we collect 82 questions that require document-level reasoning to answer. Our evaluation benchmark, although small, highlights an often unheeded problem in open-domain QA.

Our contributions can be summarized as:

1. Providing an in-depth analysis of three widely adopted open-domain QA benchmarks to identify questions for which document-level evidence is critical,
2. Introducing a new challenging benchmark of 82 questions, curated from the existing benchmarks, to highlight the importance of document-level reasoning where current models often fail.

5.2 Related Work

Retrieval in open-domain QA. In the deep learning era, open-domain QA pipelines are streamlined to a retriever plus a reader. DrQA [26], leverages a TF-IDF retriever at document-level. Subsequent works [33, 184] add a re-ranking step that recalibrates retrieval scores for paragraphs or sentences, derived from retrieved documents. Other works [100, 138] develop a model to re-rank paragraphs of retrieved docu-

ments. Yang *et al.* [194] study the granularity level of retrieval and find that retrieval at paragraph-level yields best results. Multi-passage BERT [185] suggests that fixed-length overlapping passages work best for retrieval. With the rise of dense retrievers [25] and retrieval-augmented models [103], recent models [101, 90, 70, 192, 94] have switched to passage retrieval outright.

Document modelling in QA. Several works [31, 202, 117, 181] in closed-domain QA leverage document structure to answer questions. In this work, we underline that a combination of two granularity levels (*i.e.* document and passage) is an effective means for open-domain QA.

Reasoning in QA. Datasets with different types of reasoning are prolific in QA: coreference resolution [40], multi-hop reasoning [195], numerical reasoning [47], causal relations [112], and spatial reasoning [133]. Our goal is in line with these datasets, but we focus on questions that require document-level information to answer in an open-domain setting.

5.3 Document-level Reasoning QA Challenge

5.3.1 Setup

Datasets. We use the following three popular information-seeking QA datasets: Natural Questions-OPEN (NQ-OPEN) [101], TriviaQA (TQA) [85], and WebQuestions (WQ) [19]. The details of NQ-OPEN and TQA are provided earlier in Section 2.4.1. For the sake of completeness, we give a brief overview of the three datasets:

- **Natural Questions-open (NQ-open) [101]:** Derived from Natural Questions (NQ) [98] whose questions were curated from Google search queries.
- **TriviaQA (TQA) [85]:** TQA questions are trivia questions that were mined from quiz-league websites.

- **WebQuestions (WQ) [19]**: Consisting of 2,032 questions, this dataset was collected for question answering over knowledge bases. In WQ, questions are obtained from Google Suggest API and the answers are entities whose corresponding Freebase IDs were annotated. In open-domain QA, however, the dataset is free of Freebase references and answers are stored as plain text [26].

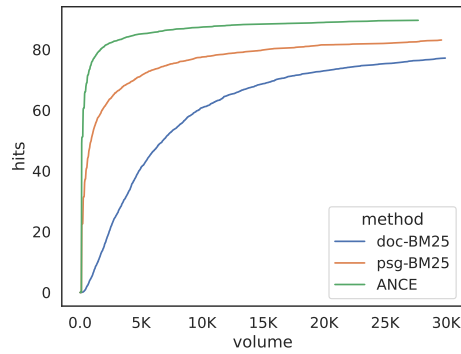
Retrieval Models. BM25 is a widely employed sparse retriever for open-domain QA, which treats text as a bag of words. We employ BM25 for both passage retrieval and document retrieval. For dense retrieval, we adopt ANCE [192], a prominent and well tested dense retriever for open-domain QA, whose trained model checkpoints are publicly available. We use three retrievers throughout the paper: ANCE and BM25 for passage retrieval, and BM25 for document retrieval.

5.3.2 Document retrieval vs. Passage retrieval

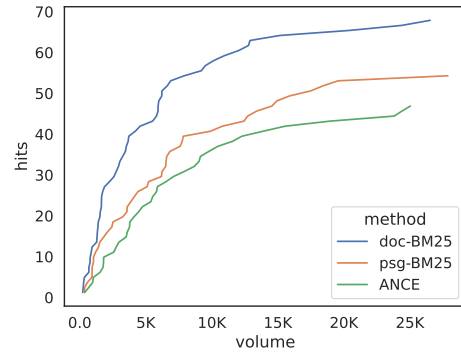
To understand when document-level reasoning is appropriate, we first need to compare the output of document retrieval against passage retrieval results. However, the ranking of candidate documents is not directly comparable to the ranking of candidate passages because of disparity in their granularity levels. To overcome the comparison problem, we compute text volume — the minimum number of terms that must be processed by a reader to find an answer in the retrieved results — to equalize the two heterogeneous granularity levels. More specifically, given that each document/-passage is a sequence of terms, we accumulate the number of terms from the top of the retrieved list until an answer is found. We measure hits ratio (*hits@vol*), the percentage of questions for which an answer document/passage is found, with respect to the text volume.

The left plots of Figure 5.2 illustrates hit ratios for the three retrievers vs. text volume. Both passage retrievers outperform document retrieval by a high margin, corroborating the consensus in the community that passage retrieval is effective on

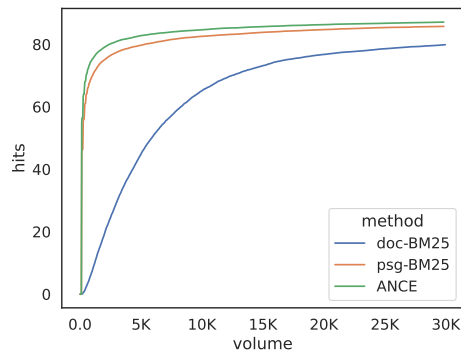
these benchmarks.



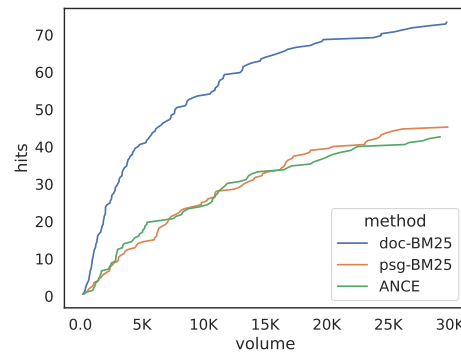
(a) NQ-OPEN (Full)



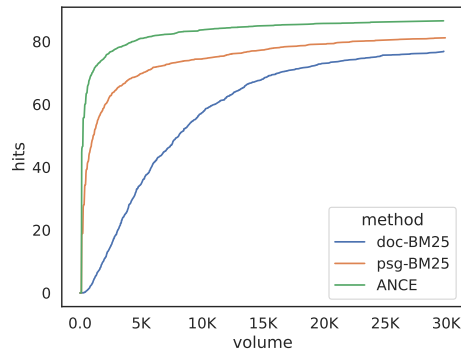
(b) NQ-OPEN (Doc wins)



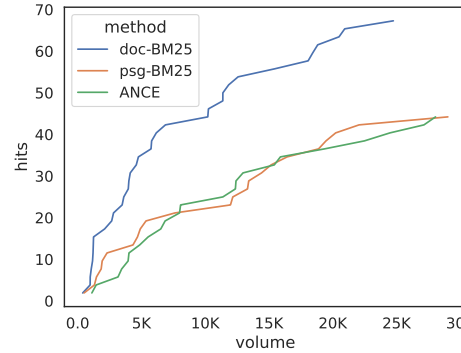
(c) TQA (Full)



(d) TQA (Doc wins)



(e) WQ (Full)



(f) WQ (Doc wins)

Figure 5.2: Hits ratio vs. text volume (left), and for the subset that document retrieval performs better (right). Although both passage retrievers outperform document retrieval by a high margin on the full dataset, document retrieval significantly outperforms both passage retrievers on the selected questions.

5.3.3 Data Collection

Based on the previous experiment, we aim to identify questions for which document retrieval surpasses passage retrieval. Table 5.1 summarizes the statistics of such questions on the three datasets. *Psg-Oracle* indicates the best passage retriever out of the two, which serves as an oracle that is aware of the better passage retriever prior to retrieval. We consider the oracle here to estimate an upper bound for passage retrieval and to make the comparison with document retrieval more robust. In total, 325 questions (4.5%), passage retrieval fails even after retrieving the same volume of text at which a naive document retrieval succeeds.

Table 5.1: Number of questions for which document retrieval surpasses passage retrieval

Dataset	#Questions that Doc wins vs.		
	Psg-BM25	ANCE	Psg-Oracle
NQ-OPEN	462 (12.8%)	177 (4.9%)	81 (2.2%)
TQA	595 (5.3%)	551 (4.9%)	192 (1.7%)
WQ	245 (12.1%)	113 (5.6%)	52 (2.6%)
<i>Total</i>	1,302	841	325

We also plot *hits@vol* varying text volume only on the selected questions, depicted in the right plots of Figure 5.2. Document retrieval surpasses both passage retrievers by a significant margin on all datasets. Interestingly, ANCE struggles most on these questions. It achieves parity with BM25 on TQA and WQ, while falling behind on NQ-OPEN.

Next, we manually audit the detected questions to shed light on:

1. What causes passage retrievers to fail?
2. Is document-level evidence really necessary to answer these questions?

To this end, we conducted a human study, done by me, to identify failure modes of passage retrieval in these questions.

Annotation Protocol. Our human annotation ensures: (1) whether top passages are sufficient for answering the question, and (2) whether the retrieved documents legitimately answer the question. Thus, the annotation procedure was done in two steps. First, for each question, the annotator checked top-2 documents, returned by the document retriever, that contain an official answer. Then, for each passage retrieval model, the annotator inspected top-5 passages. When the question was annotated as unanswerable, top-2 passages containing an official answer were also examined, if exist and not already among top-5 passages. This additional step checks whether passage retrieval correctly finds the answer or not. To select a question as a candidate for our benchmark, the annotator carefully scanned the documents to ensure document-level information including the core topic of the document, and/or the document structure is required to answer the question. This procedure took around 2 minutes per question on average.

Our manual assessment reveal three types of failure modes, showcased in Figure 5.3:

- (A) **Question-related problems:** Questions that are impossible to answer given the knowledge source [10], or are ambiguous that cannot be answered without further clarification [132].
- (B) **Answer-related problems:** Questions for which annotated answers are incorrect or miss other variations of answers that are acceptable.
- (C) **Lack of document-level understanding:** Questions that require document-level reasoning in order to determine correct answers from the knowledge source. We identify two main reasons why document-level evidence is crucial for these questions: (i) a lack of ample context, or (ii) ignoring the document narrative.

These failure modes are mutually exclusive as we did not encounter a question with both question-related issues and answer-related issues. Examples are provided in

Table 5.2.

For the majority of the questions — 75% of the audited questions — data quality problems hinder passage retrievers to find answer passages. Despite these data quality issues, we find that for nearly 25% of the questions — 82 questions in total — document-level cues are critical. These clues include an understanding of the core topic of documents or of the document structure. We present these questions as our document-level evaluation benchmark.

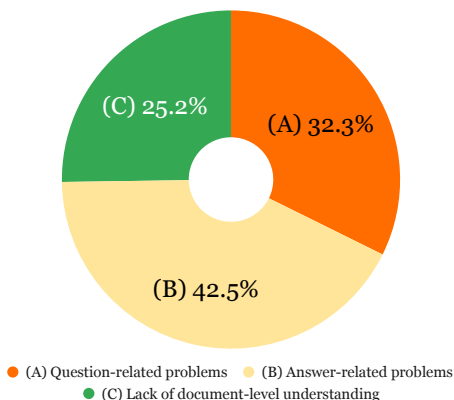


Figure 5.3: Manual inspection of 325 questions where document retrieval is superior to passage retrieval over all three open-domain QA benchmarks.

5.4 Experiments

In this section, we seek to answer the following questions to empirically show that our benchmark provides a challenge requiring document-level cues:

1. Do questions provide clues whether document-level reasoning is required?
2. What is the end-to-end performance of open-domain QA models on our benchmark?
3. Does increasing passage length provide adequate context?
4. Is injecting additional context into passages helpful?

Table 5.2: The breakdown of passage retrieval failures.

Failure Modes

(A) Question-related problems: Questions that are impossible to answer [10] given the knowledge source, or are ambiguous that cannot be answered without further clarification [132].

The lyric ‘Always sunny in a rich man’s world’, is from which song? **Unanswerable**

when did last podcast on the left start? **Unanswerable**

where will the first round of march madness be played? **Ambiguous** due to a lack of clarity in the competition year and the gender of the competition.

when did ford change the f150 body style? **Ambiguous** because Ford F-150 has been manufactured in several generations, each of which went through body changes at different times.

(B) Answer-related problems: Questions for which annotated answers are incorrect or miss other variations of answers that are acceptable.

how many times has psg won champions league? The official answer is **46**, but the actual answer is **0**.

what is the first line of an http request to the server? Official answers are **a request line** or **the status line**, but the answer is **the first line of the response**.

where did the battle of bonhomme richard take place? The official answer is **near Flamborough Head**, yet **the North Sea off the coast of Yorkshire** or **Flamborough Head, Yorkshire** are plausible too.

In the mid 1990s what major fossil discovery was made in Liaoning, China? While the official answer is **Well-preserved fossils of feathered dinosaurs**, **feathered dinosaur fossils** is also acceptable.

(C) Lack of document-level understanding: Questions that require document-level reasoning in order to determine their answers from the knowledge source.

when was the last time the boston red sox pitched a no-hitter? (NQ-OPEN)

In which film did teacher John Keating break all the rules? (TQA)

what high school did maya angelou go to? (WQ)

5.5 Experimental Setup

For retrieval, our knowledge source is Wikipedia articles, corresponding to the snapshot of 20-Dec-2018, following [101, 90]. We used Wikipedia passages, provided by DPR [90]. Specifically, Wikipedia articles were split into non-overlapping passages of 100 words [185] along with the article title that is concatenated to the start of each passage. For sparse retrieval, we construct the inverted index using Pyserini [111]

and for dense retrieval, we use pre-encoded index files from Pygaggle².

Tuning BM25. For passage retrieval, we use $k_1 = 0.9$ and $b = 0.4$, as reported in DPR. For document retrieval, k_1 and b were tuned on the dev set of each dataset separately. We bootstrap k_1 and b by repeatedly resampling from $[0, 3]$ and $[0, 1]$ (ranges are taken from [178]), 100 times with replacement. We take the best k_1 and b pairs; on NQ-OPEN: $k_1 = 2.5$ and $b = 0.3$, on TQA: $k_1 = 1.5$ and $b = 0.2$, and on WQ: $k_1 = 2.9$ and $b = 0.3$.

5.5.1 Predicting Granularity Level of Retrieval

First, we study whether document-level cues can be identified using only questions. To this end, we build a binary classifier that is able to accurately predicting the granularity level of retrieval (*i.e.*, passage-level or document-level). Such a classifier is reminiscent of the unanswerability prediction via only the question [10] that achieves an accuracy of 73%. Similarly, we train a classifier that takes a question as input and predicts whether retrieval should be done at document-level or not.

The training data is constructed by computing text volume of BM25, explained in §5.3.2, for both passage retrieval vol_{psg} and document retrieval vol_{doc} on the training set of NQ-OPEN. The label of each question is $\text{argmin}(vol_{\text{psg}}, vol_{\text{doc}})$. The dataset statistics, constructed using this method, is reported in Table 5.3. We fine-tuned RoBERTa_{base} [118] on this dataset for 5 epochs (learning rate= $1.5e^{-5}$ with a linear decay and a warmup ratio of 0.1), and with a weighted cross entropy loss³ to account for data imbalance. Our classifier achieves an accuracy of 65.7% (AUC=0.665, and recall=58.7%) on the test set.

Next, we plug in the classifier to an open-domain QA pipeline with BM25 as the retriever and FiD [81] as the reader. This model achieves an exact match (EM) ac-

²<https://github.com/castorini/pygaggle/>

³We adopted the “balanced” heuristic from the scikit-learn package for computing class weights: https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

Table 5.3: Dataset Statistic, constructed from NQ-OPEN, for predicting the granularity level of retrieval.

Data Split	Size	#Doc-labelled instances
Train	69,896	9,848 (14.1%)
Dev	8,757	1,125 (4.9%)
Test	3,610	858 (5.6%)

curacy of 39.7% and 8.8% on the full NQ-OPEN, and our benchmark, respectively. The performance of the pipeline deteriorates, compared to when the classifier is not used (Table 5.4), which indicates that our classifier is not accurate enough in predicting the right granularity level of the retrieval. Hence, unlike unanswerability [10], using only the question to predict the granularity level of retrieval is not a useful tool for an open-domain QA pipeline. Moreover, these results indicate that the need for document-level reasoning is not a characteristic of the question alone. This also supports our observation in the human study that document-level reasoning is a byproduct of the corpus rather than the questions.

5.5.2 End-to-End Results

To measure the end-to-end performance, we pair our retrievers with the well-established FiD reader [81], whose trained checkpoint is publicly available. For document retrieval, retrieved documents are naively split into passages as FiD accepts only passages. Note that this approach is not efficient and serves merely as a baseline. We restrict the number of passages that are fed to the reader to 100, analogous to previous work [81, 90]. This restriction indeed puts document retrieval at disadvantage since some parts of documents may be cut off. As showcased in Table 5.4, document retrieval with our naive approach substantially underperforms on the full dataset, whereas it leads both passage retrievers on our benchmark. This result highlights that document-level information is central to answer the questions in our benchmark

that even our naive approach surpasses the full-fledged passage-based pipelines.

Table 5.4: Exact-match accuracy of our retrievers, paired with FiD [81], on NQ-OPEN.

Pipeline	Full Dataset	Our benchmark
Psg-BM25 + FiD	41.4	8.6
ANCE + FiD	46.6	6.9
Doc-BM25 + FiD	33.5	12.1

5.5.3 Varying Passage Length

One hypothesis is that increasing the passage length can be helpful when passages are not long enough to reflect the document discourse. To investigate this, we vary passage length within {50, 100, 200, 500, 1000} and perform BM25 retrieval for each passage length. To this end, we construct a separate index for each passage length and tune BM25 parameters as explained in §5.5. Then, we retrieve passages using BM25 over each index and measure hits ratio at volume 10K. The results are visualized in Figure 5.4 for the full NQ-OPEN as well as our evaluation benchmark. Even though the performance declines overall with longer passage lengths, the hits ratio actually increases on our document-level benchmark. The performance of document retrieval remains almost intact for both cases. These results indicate that more context is indeed required to locate plausible candidates on our benchmark.

5.5.4 Enrich Passages with Additional Context

In light of our findings in Section 5.5.3, we investigate whether enriching passages with additional context is helpful or not. For this purpose, we leverage the leading section of a document since according to the Wikipedia writing guidelines⁴, it serves as a “summary of the most important points” of the document and “should stand on its own as a concise overview of the article’s topic.”

⁴https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

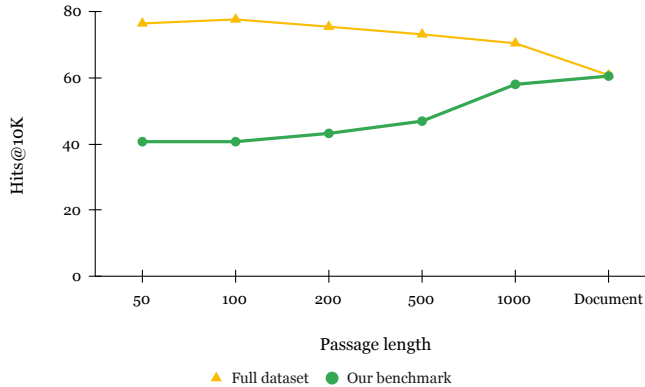


Figure 5.4: Hits ratio at volume 10K for various passage lengths on NQ-OPEN and our benchmark.

We take the first passage of documents as an estimate of the leading section and prepend it to each passage in the document. We retrieve the enriched passages using BM25 whose parameters are tuned following the procedure explained in Section 5.5. The retrieval performance, hits ratio at volume 10K, slightly drops to 76.7%, compared to that of passage retrieval — 77.6% — on NQ-OPEN. On our benchmark, enriched passage retrieval achieves 57.5%, whereas passage retrieval scores 47.5%. We further plug the enriched passage retrieval into FiD. The exact match accuracy for 100 enriched passages bumps up to 10.0%. Note that enriched passages are approximately twice longer than a passage, and they are mostly composed of repetitive content (*i.e.* the first passage of their corresponding documents). Hence, this overall result is not directly comparable with the end-to-end results, presented in Table 5.4. Yet, these findings emphasize that additional context is indeed helpful for our benchmark.

5.6 Conclusion

Passage retrieval is not always sufficient for open-domain QA models. In fact, answering some questions requires document-level reasoning. In this paper, we show that this phenomenon is largely overlooked in existing benchmarks. To this end, we

introduce a novel benchmark, carefully curated from three well-known open-domain QA datasets, that consists of 82 such questions. We hope our benchmark spurs the development of document-level open-domain QA models.

Chapter 6

Conclusion and Future Work

This thesis culminates with drawing the central conclusions from the previous chapters. To this end, I first recap the key contributions made towards *building robust and scalable QA systems suitable for real-world applications*, my primary goal in this thesis. Then, I take stock of the findings and contributions to discuss potential future research directions.

6.1 Summary of Contributions

The present thesis set out to alleviate the adoption of open-domain QA models, capable of answering information-seeking questions over a knowledge source, in applied areas. I argued that achieving this goal demands the following four key desiderata, stated in Section 1.2: (D1) ability to effectively acquire knowledge from text, (D2) maintaining robustness under distribution shifts during test time, (D3) ability to scale up for large volume of data, and (D4) reliability in evaluation. In this section, I discuss the contributions of my research and their connections to these desiderata. The core of the work I presented in this thesis can be viewed through the lenses of modelling, data, and evaluation, a vital triad for building clear-cut real-world systems.

I first focused on retrieval in open-domain QA systems (Chapter 2). Traditional sparse retrieval has long been the go-to approach in open-domain QA until recently. However, they are not designed for QA where textual cues matter most. This problem

concerns a lack of knowledge acquisition methods that can incorporate some useful features into the retrieval. To overcome this problem, I presented a simple retrieval model to endow sparse retrievers with word order, the most salient syntactic cue. Knowing what words precede and/or succeed a word in a sentence, also referred to as *local word order*, strengthens the likelihood of localized context in documents. The proposed model is shown to be a strong baseline for dense retrieval models. Local word order-aware retrieval complements dense models when fused with them, thus suggesting that the two models flounder on different test cases. Unlike dense models (DPR [90], ANCE [192], *inter alia*) that perform poorly on out-of-distribution data [177, 143], sparse models remain robust under distribution shift.

In Chapter 3, I proceeded to work on another crucial component in open-domain QA systems, machine readers whose objective is to precisely pinpoint the answer. I focused on a well-documented issue in this regard. QA models are heavily susceptible to domain shift [130, 88, 15]. I explored the impact of DA to overcome this problem because DA is previously shown effective for in-domain tests [4]. However, the size of augmented data that directly impacts the training speed is never studied. Using DA blindly impairs the scalability of training on large-scale datasets. Since this issue is not limited to QA, I shifted gears to a wide range of NLU tasks including reading comprehension (*i.e.* answering a question, given a passage). I presented an algorithm, namely Glitter, that adaptively selects a subset of worst-case augmented samples with a maximal loss. Glitter is flexible in the choice of the selection criterion; it can be as simple as a cross entropy loss. More importantly, Glitter is sample-efficient in that it selects only a portion of augmented data. Also, Glitter can be coupled with any DA method, making it a universal framework. I showed that Glitter does not compromise accuracy, but it is significantly faster to train, compared to naively using all augmented data.

In Chapter 4, in line with the goal of knowledge acquisition methods, I studied the task of generating viable meta information from bare-bones text, a path towards

building context-rich text collections. In particular, I developed an algorithm to map location mentions, called toponyms, to their corresponding spatial coordinates without a reliance on labelled data. The proposed algorithm is based on two intuitive hypotheses that (i) the geographic scope of the document is a strong signal to locate toponyms, and (ii) nearby toponyms tend to be linked with one another. The two hypotheses were characterized by a probabilistic model. However, one missing element in the algorithm was to leverage spatial hierarchies (*e.g.* containment and sibling relationships). For this purpose, I considered the minimality hypothesis of toponyms. Specifically, toponyms in a document are often in the spatial proximity of each other. The minimality property is modelled via the classical conflict-free covering problem. All these hypotheses combined form a strong unsupervised model for the task that surpasses supervised models on previously unseen toponyms.

Finally, Chapter 5 looked into the reliability of evaluation benchmarks in open-domain QA. When comparing document retrieval with passage retrieval on existing well-known benchmarks, the results suggest that passage retrieval outperforms document retrieval by a high margin. These results have inspired many passage retrieval models in the community insofar as document retrieval has gone virtually extinct in recent open-domain QA models. Admittedly, the use of passage retrieval has a strong precedent in this task [35] and is based on the assumption that the knowledge source is substantial enough that the answer can be found somewhere in a localized context. Nonetheless, this assumption does not hold in practice for Wikipedia and other specialized domains. Such observations motivated me to investigate questions for which document retrieval performs better than passage retrieval. To this end, I conducted a thorough manual analysis to find the real reason of the superiority of document retrieval. The human study identified a small set of questions that are impossible to be answered using passage retrieval because some kind of document-level evidence is necessary for answering them. I curated these questions as a document-level reasoning benchmark for open-domain QA to spur the development of QA models with

document-level understanding.

6.2 Future Work

I strongly believe there is still a long way to go to accomplish genuine human-level open-domain QA. The discussion in Section 6.1 leaves ample avenues for future work. In this section, I discuss the open problems based on the insights I gained doing this thesis.

6.2.1 Considerations for using models in the wild

I touched upon four key requirements for deploying systems in real-world applications. Nonetheless, systems that deal with real life problems even though successful on research benchmarks are facing a myriad of other concerns that I envision as challenging future directions: *interpretability*, and *fairness*.

Interpretability: The IR and NLP communities are used to model transparency because the behaviour of classical statistical models were often predictable and controllable. In contrast, in the deep learning era, neural models lack such transparency. We need tools to provide explanation why a model is producing a particular output. Moreover, we should be able to control the behaviour of a model because it provides a means to diagnose and fix issues in models.

Fairness: Models should fairly treat diverse perspectives. Recent colossal PLMs are shown to reflect biases and perpetuate stereotypes [62, 17]. How to surmount biases, present in data, is an indispensable open problem. Another question is what modelling decisions are needed to overcome biases. PLMs are nowadays the dominant backbone of QA models, denoting that they propagate their biases into downstream tasks. Another pivotal aspect of this problem is heeding the social contexts and target demographics when models are considered to be used in practice.

6.2.2 Scalability and Complex Reasoning

Answering questions requires a host of various reasoning skills including numerical reasoning [5], discrete processing [47], coreferential understanding [40], and multi-hop reasoning [195]. In fact, QA datasets can be categorized based on reasoning capabilities, but they are mostly studied in a closed-domain setting where a context is given. A potential future direction can be studying complex questions in an open-domain mode where we have only access to a massive knowledge source. Multi-hop QA is a leading example in this avenue where multiple supporting evidence should be retrieved from a knowledge source for finding an answer. Making complex reasoning QA tasks open introduces scalability as another challenging dimension into the problem and may give rise to novel retrieval models.

6.2.3 Sample Efficiency for Dense Retrieval

Sample efficiency or learning from fewer examples allows us to build scalable models on large-scale datasets. In IR, training end-to-end dense retrieval models exhibits two limitations. First, they need a large number of queries, comprising positive and negative examples. Second, these models should frequently re-index the data during training [70, 170]. These problems entail the need for vast computational resources that amount to larger carbon footprint and inhibit the wide adoption of these models. Strategies to make dense retrievers sample-efficient would make them easier to build and more accessible.

6.2.4 Data-centric Analysis of QA Datasets

Data-centric practices aim at systematically engineering data to build successful models. The increasing interest in data-centric techniques suggests that we need to look for sustainable ways to analyze existing datasets. Our manual assessment of three popular open-domain QA benchmarks in Section 5.3.3 uncovers a troubling number of data quality issues. A careful inspection of these errors and the steps to circumvent

them are potential directions that can lead to the creation of high-quality datasets in the future.

Bibliography

- [1] M. D. Adelfio and H. Samet, “Geowhiz: Toponym resolution using common categories,” in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL’13, Orlando, Florida: ACM, 2013, pp. 532–535, ISBN: 978-1-4503-2521-9. DOI: 10.1145/2525314.2525321. [Online]. Available: <http://doi.acm.org/10.1145/2525314.2525321>.
- [2] D. Ahlers, “Assessment of the accuracy of geonames gazetteer data,” in *Proceedings of the 7th Workshop on Geographic Information Retrieval*, 2013, pp. 74–81.
- [3] J. Ainslie, S. Ontanon, C. Alberti, V. Cvicek, Z. Fisher, P. Pham, A. Ravula, S. Sanghai, Q. Wang, and L. Yang, “ETC: Encoding long and structured inputs in transformers,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 268–284. DOI: 10.18653/v1/2020.emnlp-main.19. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.19>.
- [4] C. Alberti, D. Andor, E. Pitler, J. Devlin, and M. Collins, “Synthetic QA corpora generation with roundtrip consistency,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6168–6173. DOI: 10.18653/v1/P19-1620. [Online]. Available: <https://aclanthology.org/P19-1620>.
- [5] A. Amini, S. Gabriel, S. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi, “MathQA: Towards interpretable math word problem solving with operation-based formalisms,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2357–2367. DOI: 10.18653/v1/N19-1245. [Online]. Available: <https://aclanthology.org/N19-1245>.
- [6] E. Amitay, N. Har’El, R. Sivan, and A. Soffer, “Web-a-where: Geotagging web content,” in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’04, Sheffield, United Kingdom: ACM, 2004, pp. 273–280, ISBN: 1-58113-881-4.

- DOI: 10.1145/1008992.1009040. [Online]. Available: <http://doi.acm.org/10.1145/1008992.1009040>.
- [7] G. Andogah, G. Bouma, and J. Nerbonne, “Every document has a geographical scope,” *Data Knowl. Eng.*, vol. 81-82, pp. 1–20, Nov. 2012, ISSN: 0169-023X. DOI: 10.1016/j.datak.2012.07.002. [Online]. Available: <http://dx.doi.org/10.1016/j.datak.2012.07.002>.
- [8] M. C. Ardanuy and C. Sporleder, “Toponym disambiguation in historical documents using semantic and geographic features,” in *Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage*, ser. DATeCH2017, Gttingen, Germany: ACM, 2017, pp. 175–180, ISBN: 978-1-4503-5265-9. DOI: 10.1145/3078081.3078099. [Online]. Available: <http://doi.acm.org/10.1145/3078081.3078099>.
- [9] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant risk minimization,” *arXiv preprint arXiv:1907.02893*, 2019.
- [10] A. Asai and E. Choi, “Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 1492–1504. DOI: 10.18653/v1/2021.acl-long.118. [Online]. Available: <https://aclanthology.org/2021.acl-long.118>.
- [11] A. Asai, M. Gardner, and H. Hajishirzi, “Evidentiality-guided generation for knowledge-intensive NLP tasks,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 2226–2243. [Online]. Available: <https://aclanthology.org/2022.naacl-main.162>.
- [12] A. Asai and H. Hajishirzi, “Logic-guided data augmentation and regularization for consistent question answering,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 5642–5650. DOI: 10.18653/v1/2020.acl-main.499. [Online]. Available: <https://aclanthology.org/2020.acl-main.499>.
- [13] A. Asai, K. Hashimoto, H. Hajishirzi, R. Socher, and C. Xiong, “Learning to retrieve reasoning paths over wikipedia graph for question answering,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=SJgVHkrYDH>.
- [14] M. Banko, E. Brill, S. Dumais, and J. Lin, “AskMSR: Question answering using the worldwide web,” in *Proceedings of 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002, pp. 7–9.

- [15] M. Bartolo, T. Thrush, R. Jia, S. Riedel, P. Stenetorp, and D. Kiela, “Improving question answering model robustness with synthetic adversarial data generation,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8830–8848. DOI: 10.18653/v1/2021.emnlp-main.696. [Online]. Available: <https://aclanthology.org/2021.emnlp-main.696>.
- [16] E. M. Bender and B. Friedman, “Data statements for natural language processing: Toward mitigating system bias and enabling better science,” *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018. DOI: 10.1162/tacl.a.00041. [Online]. Available: <https://aclanthology.org/Q18-1041>.
- [17] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2021, pp. 610–623. DOI: 10.1145/3442188.3445922.
- [18] M. Bendersky, D. Metzler, and W. B. Croft, “Parameterized concept weighting in verbose queries,” in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Beijing, China: Association for Computing Machinery, 2011, pp. 605–614. DOI: 10.1145/2009916.2009998.
- [19] J. Berant, A. Chou, R. Frostig, and P. Liang, “Semantic parsing on Freebase from question-answer pairs,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1533–1544. [Online]. Available: <https://aclanthology.org/D13-1160>.
- [20] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, “Bridging the lexical chasm: Statistical approaches to answer-finding,” in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athens, Greece: Association for Computing Machinery, 2000, pp. 192–199. DOI: 10.1145/345508.345576.
- [21] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. De Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. Rae, E. Elsen, and L. Sifre, “Improving language models by retrieving from trillions of tokens,” in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 17–23 Jul 2022, pp. 2206–2240. [Online]. Available: <https://proceedings.mlr.press/v162/borgeaud22a.html>.

- [22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- [23] C. Buciluă, R. Caruana, and A. Niculescu-Mizil, “Model compression,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 535–541.
- [24] D. Buscaldi, “Approaches to disambiguating toponyms,” *SIGSPATIAL Special*, vol. 3, no. 2, pp. 16–19, Jul. 2011, ISSN: 1946-7729. DOI: 10.1145/2047296.2047300. [Online]. Available: <http://doi.acm.org/10.1145/2047296.2047300>.
- [25] W.-C. Chang, F. X. Yu, Y.-W. Chang, Y. Yang, and S. Kumar, “Pre-training tasks for embedding-based large-scale retrieval,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=rkg-mA4FDr>.
- [26] D. Chen, A. Fisch, J. Weston, and A. Bordes, “Reading Wikipedia to answer open-domain questions,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1870–1879. DOI: 10.18653/v1/P17-1171. [Online]. Available: <https://aclanthology.org/P17-1171>.
- [27] J. Chen, D. Shen, W. Chen, and D. Yang, “HiddenCut: Simple data augmentation for natural language understanding with better generalizability,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 4380–4390. DOI: 10.18653/v1/2021.acl-long.338. [Online]. Available: <https://aclanthology.org/2021.acl-long.338>.
- [28] Y. Chen and J. Joo, “Understanding and mitigating annotation bias in facial expression recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14 980–14 991.
- [29] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: A content-based approach to geo-locating twitter users,” in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’10, Toronto, ON, Canada: ACM, 2010, pp. 759–768, ISBN: 978-1-4503-0099-5. DOI: 10.1145/1871437.1871535. [Online]. Available: <http://doi.acm.org/10.1145/1871437.1871535>.

- [30] J. H. Cho and B. Hariharan, “On the efficacy of knowledge distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4794–4802.
- [31] E. Choi, D. Hewlett, J. Uszkoreit, I. Polosukhin, A. Lacoste, and J. Berant, “Coarse-to-fine question answering for long documents,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 209–220. DOI: 10.18653/v1/P17-1020. [Online]. Available: <https://aclanthology.org/P17-1020>.
- [32] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, “PaLM: Scaling language modeling with pathways,” *arXiv preprint arXiv:2204.02311*, 2022.
- [33] C. Clark and M. Gardner, “Simple and effective multi-paragraph reading comprehension,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 845–855. DOI: 10.18653/v1/P18-1078. [Online]. Available: <https://aclanthology.org/P18-1078>.
- [34] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, “Think you have solved question answering? try ARC, the AI2 reasoning challenge,” *arXiv preprint arXiv:1803.05457*, 2018.
- [35] C. L. Clarke, G. V. Cormack, and T. R. Lynam, “Exploiting redundancy in question answering,” in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 358–365.
- [36] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “AutoAugment: Learning augmentation policies from data,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [37] Z. Dai and J. Callan, “Deeper text understanding for IR with contextual neural language modeling,” in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Paris, France: Association for Computing Machinery, 2019, pp. 985–988. DOI: 10.1145/3331184.3331303.
- [38] Z. Dai, C. Xiong, J. Callan, and Z. Liu, “Convolutional neural networks for soft-matching n-grams in ad-hoc search,” in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, pp. 126–134. DOI: 10.1145/3159652.3159659.
- [39] E. K. Dang, R. W. Luk, and J. Allan, “Beyond bag-of-words: Bigram-enhanced context-dependent term weights,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 6, pp. 1134–1148, 2014.

- [40] P. Dasigi, N. F. Liu, A. Marasović, N. A. Smith, and M. Gardner, “Quoref: A reading comprehension dataset with questions requiring coreferential reasoning,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5925–5932. DOI: 10.18653/v1/D19-1606. [Online]. Available: <https://aclanthology.org/D19-1606>.
- [41] G. DeLozier, J. Baldridge, and L. London, “Gazetteer-independent toponym resolution using geographic word profiles,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI’15, Austin, Texas: AAAI Press, 2015, pp. 2382–2388, ISBN: 0-262-51129-0. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2886521.2886652>.
- [42] D. Demszky, K. Guu, and P. Liang, “Transforming question answering datasets into natural language inference datasets,” *arXiv preprint arXiv:1809.02922*, 2018.
- [43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: <https://aclanthology.org/N19-1423>.
- [44] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” *arXiv preprint arXiv:1708.04552*, 2017.
- [45] P. Domingos, “A few useful things to know about machine learning,” *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012, ISSN: 0001-0782. DOI: 10.1145/2347736.2347755. [Online]. Available: <http://doi.acm.org/10.1145/2347736.2347755>.
- [46] J. Du, E. Grave, B. Gunel, V. Chaudhary, O. Celebi, M. Auli, V. Stoyanov, and A. Conneau, “Self-training improves pre-training for natural language understanding,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 5408–5418. DOI: 10.18653/v1/2021.naacl-main.426. [Online]. Available: <https://aclanthology.org/2021.naacl-main.426>.
- [47] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, “DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2368–2378. DOI: 10.18653/v1/N19-1246. [Online]. Available: <https://aclanthology.org/N19-1246>.

- [48] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng, “Web question answering: Is more always better?” In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland: Association for Computing Machinery, 2002, pp. 291–298. DOI: 10.1145/564376.564428.
- [49] S. Edunov, M. Ott, M. Auli, and D. Grangier, “Understanding back-translation at scale,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 489–500. DOI: 10.18653/v1/D18-1045. [Online]. Available: <https://aclanthology.org/D18-1045>.
- [50] C. Eickhoff, A. P. de Vries, and T. Hofmann, “Modelling term dependence with copulas,” in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile: Association for Computing Machinery, 2015, pp. 783–786. DOI: 10.1145/2766462.2767831.
- [51] J. Eisenstein, “Phonological factors in social media writing,” in *Proceedings of the Workshop on Language Analysis in Social Media*, Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 11–19. [Online]. Available: <http://www.aclweb.org/anthology/W13-1102>.
- [52] Y. Fan, J. Guo, X. Ma, R. Zhang, Y. Lan, and X. Cheng, “A linguistic study on relevance modeling in information retrieval,” in *Proceedings of the Web Conference 2021*, ser. WWW ’21, Association for Computing Machinery, 2021, pp. 1053–1064. DOI: 10.1145/3442381.3450009.
- [53] F. Farnia and D. Tse, “A minimax approach to supervised learning,” *Advances in Neural Information Processing Systems*, vol. 29, pp. 4240–4248, 2016. [Online]. Available: <https://dl.acm.org/doi/pdf/10.5555/3157382.3157571>.
- [54] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, “A survey of data augmentation approaches for NLP,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, Aug. 2021, pp. 968–988. DOI: 10.18653/v1/2021.findings-acl.84. [Online]. Available: <https://aclanthology.org/2021.findings-acl.84>.
- [55] J. R. Finkel, T. Grenager, and C. Manning, “Incorporating non-local information into information extraction systems by gibbs sampling,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL ’05, Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 363–370. DOI: 10.3115/1219840.1219885. [Online]. Available: <https://doi.org/10.3115/1219840.1219885>.
- [56] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza, “On the accuracy of hyper-local geotagging of social media content,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*,

- ser. WSDM '15, Shanghai, China: ACM, 2015, pp. 127–136, ISBN: 978-1-4503-3317-7. DOI: 10.1145/2684822.2685296. [Online]. Available: <http://doi.acm.org/10.1145/2684822.2685296>.
- [57] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, “Born again neural networks,” in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, 2018, pp. 1607–1616. [Online]. Available: <https://proceedings.mlr.press/v80/furlanello18a.html>.
- [58] O.-E. Ganea, M. Ganea, A. Lucchi, C. Eickhoff, and T. Hofmann, “Probabilistic bag-of-hyperlinks model for entity linking,” in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16, Montrécal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 927–938, ISBN: 978-1-4503-4143-1. DOI: 10.1145/2872427.2882988. [Online]. Available: <https://doi.org/10.1145/2872427.2882988>.
- [59] L. Gao, Z. Dai, and J. Callan, “COIL: Revisit exact lexical match in information retrieval with contextualized inverted list,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 3030–3042. DOI: 10.18653/v1/2021.naacl-main.241. [Online]. Available: <https://aclanthology.org/2021.naacl-main.241>.
- [60] E. Garbin and I. Mani, “Disambiguating toponyms in news,” in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 2005, pp. 363–370. [Online]. Available: <http://www.aclweb.org/anthology/H/H05/H05-1046>.
- [61] M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, N. Gupta, H. Hajishirzi, G. Ilharco, D. Khashabi, K. Lin, J. Liu, N. F. Liu, P. Mulcaire, Q. Ning, S. Singh, N. A. Smith, S. Subramanian, R. Tsarfaty, E. Wallace, A. Zhang, and B. Zhou, “Evaluating models’ local decision boundaries via contrast sets,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1307–1323. DOI: 10.18653/v1/2020.findings-emnlp.117. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.117>.
- [62] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, “RealToxicityPrompts: Evaluating neural toxic degeneration in language models,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 3356–3369. DOI: 10.18653/v1/2020.findings-emnlp.301. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.301>.

- [63] R. Geirhos, J. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, “Shortcut learning in deep neural networks,” *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020. DOI: 10.1038/s42256-020-00257-z.
- [64] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness,” in *International Conference on Learning Representations*, 2018.
- [65] J. Gelernter, G. Ganesh, H. Krishnakumar, and W. Zhang, “Automatic gazetteer enrichment with user-geocoded data,” in *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, ACM, 2013, pp. 87–94.
- [66] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations*, 2014.
- [67] B. F. Green Jr, A. K. Wolf, C. Chomsky, and K. Laughery, “Baseball: An automatic question-answerer,” in *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, Association for Computing Machinery, 1961. DOI: 10.1145/1460690.1460714.
- [68] H. Guo, Y. Mao, and R. Zhang, “Augmenting data with mixup for sentence classification: An empirical study,” *arXiv preprint arXiv:1905.08941*, 2019.
- [69] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, “A deep relevance matching model for ad-hoc retrieval,” in *Proceedings of the 25th ACM international on conference on information and knowledge management*, Indianapolis, Indiana, USA: Association for Computing Machinery, 2016, pp. 55–64. DOI: 10.1145/2983323.2983769.
- [70] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Retrieval augmented language model pre-training,” in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, PMLR, 2020, pp. 3929–3938. [Online]. Available: <https://proceedings.mlr.press/v119/guu20a.html>.
- [71] S. Har-Peled and S. Smorodinsky, “On conflict-free coloring of points and simple regions in the plane,” in *Proceedings of the Nineteenth Annual Symposium on Computational Geometry*, ser. SCG ’03, San Diego, California, USA: ACM, 2003, pp. 114–123, ISBN: 1-58113-663-3. DOI: 10.1145/777792.777810. [Online]. Available: <http://doi.acm.org/10.1145/777792.777810>.
- [72] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, “Pretrained transformers improve out-of-distribution robustness,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 2744–2751. DOI: 10.18653/v1/2020.acl-main.244. [Online]. Available: <https://aclanthology.org/2020.acl-main.244>.

- [73] D. Hendrycks*, N. Mu*, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, “AugMix: A simple method to improve robustness and uncertainty under data shift,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=S1gmrxFvB>.
- [74] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [75] J. Hoffart, Y. Altun, and G. Weikum, “Discovering emerging entities with ambiguous names,” in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW ’14, Seoul, Korea: ACM, 2014, pp. 385–396, ISBN: 978-1-4503-2744-2. DOI: 10.1145/2566486.2568003. [Online]. Available: <http://doi.acm.org/10.1145/2566486.2568003>.
- [76] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=rygGQyrFvH>.
- [77] Z. Hu, B. Tan, R. Salakhutdinov, T. Mitchell, and E. P. Xing, “Learning data manipulation for augmentation and weighting,” *arXiv preprint arXiv:1910.12795*, 2019.
- [78] T. Huang, S. Halbe, C. Sankar, P. Amini, S. Kottur, A. Geramifard, M. Razaviyayn, and A. Beirami, “DAIR: Data augmented invariant regularization,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=PKdNRKjwL4>.
- [79] Y. Ikawa, M. Vukovic, J. Rogstadius, and A. Murakami, “Location-based insights from the social web,” in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW ’13 Companion, Rio de Janeiro, Brazil: ACM, 2013, pp. 1013–1016, ISBN: 978-1-4503-2038-2. DOI: 10.1145/2487788.2488107. [Online]. Available: <http://doi.acm.org/10.1145/2487788.2488107>.
- [80] G. Izacard and E. Grave, “Distilling knowledge from reader to retriever for question answering,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=NTEz-6wysdb>.
- [81] G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 874–880. DOI: 10.18653/v1/2021.eacl-main.74. [Online]. Available: <https://aclanthology.org/2021.eacl-main.74>.
- [82] R. Jia and P. Liang, “Adversarial examples for evaluating reading comprehension systems,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2021–2031. DOI: 10.18653/v1/D17-1215. [Online]. Available: <https://aclanthology.org/D17-1215>.

- [83] J.-Y. Jiang, M. Zhang, C. Li, M. Bendersky, N. Golbandi, and M. Najork, “Semantic text matching for long-form documents,” in *The World Wide Web Conference*, ser. WWW ’19, Association for Computing Machinery, 2019, pp. 795–806, ISBN: 9781450366748. DOI: 10.1145/3308558.3313707.
- [84] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, “TinyBERT: Distilling BERT for natural language understanding,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 4163–4174. DOI: 10.18653/v1/2020.findings-emnlp.372. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.372>.
- [85] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, “TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1601–1611. DOI: 10.18653/v1/P17-1147. [Online]. Available: <https://aclanthology.org/P17-1147>.
- [86] D. Jurafsky and J. H. Martin. (2022). *Speech and language processing* (3rd ed. draft), [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/> (visited on 01/12/2022).
- [87] E. Kamaloo, M. Rezagholizadeh, P. Passban, and A. Ghodsi, “Not far away, not so close: Sample efficient nearest neighbour data augmentation via MiniMax,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, Aug. 2021, pp. 3522–3533. DOI: 10.18653/v1/2021.findings-acl.309. [Online]. Available: <https://aclanthology.org/2021.findings-acl.309>.
- [88] A. Kamath, R. Jia, and P. Liang, “Selective question answering under domain shift,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 5684–5696. DOI: 10.18653/v1/2020.acl-main.503. [Online]. Available: <https://aclanthology.org/2020.acl-main.503>.
- [89] K. Karkkainen and J. Joo, “FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1548–1558.
- [90] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, “Dense passage retrieval for open-domain question answering,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.550>.

- [91] D. Kaushik, E. Hovy, and Z. Lipton, “Learning the difference that makes a difference with counterfactually-augmented data,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=SkLgs0NFvr>.
- [92] U. Khandelwal, H. He, P. Qi, and D. Jurafsky, “Sharp nearby, fuzzy far away: How neural language models use context,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 284–294. DOI: 10.18653/v1/P18-1027. [Online]. Available: <https://aclanthology.org/P18-1027>.
- [93] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi, “UNIFIEDQA: Crossing format boundaries with a single QA system,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1896–1907. DOI: 10.18653/v1/2020.findings-emnlp.171. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.171>.
- [94] O. Khattab, C. Potts, and M. Zaharia, “Relevance-guided supervision for OpenQA with ColBERT,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 929–944, 2021. DOI: 10.1162/tacl.a.00405. [Online]. Available: <https://aclanthology.org/2021.tacl-1.55>.
- [95] T. Khot, A. Sabharwal, and P. Clark, “SciTail: A textual entailment dataset from science question answering,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [96] S. Kobayashi, “Contextual augmentation: Data augmentation by words with paradigmatic relations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 452–457. DOI: 10.18653/v1/N18-2072. [Online]. Available: <https://aclanthology.org/N18-2072>.
- [97] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, “WILDS: A benchmark of in-the-wild distribution shifts,” in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 18–24 Jul 2021, pp. 5637–5664. [Online]. Available: <https://proceedings.mlr.press/v139/koh21a.html>.
- [98] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelleey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, “Natural questions: A benchmark for question answering research,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 452–466, 2019. DOI: 10.1162/tacl.a.00276. [Online]. Available: <https://aclanthology.org/Q19-1026>.

- [99] C. Kwok, O. Etzioni, and D. S. Weld, “Scaling question answering to the web,” *ACM Transactions on Information Systems (TOIS)*, vol. 19, no. 3, pp. 242–262, 2001.
- [100] J. Lee, S. Yun, H. Kim, M. Ko, and J. Kang, “Ranking paragraphs for improving answer recall in open-domain question answering,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 565–569. DOI: 10.18653/v1/D18-1053. [Online]. Available: <https://aclanthology.org/D18-1053>.
- [101] K. Lee, M.-W. Chang, and K. Toutanova, “Latent retrieval for weakly supervised open domain question answering,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6086–6096. DOI: 10.18653/v1/P19-1612. [Online]. Available: <https://aclanthology.org/P19-1612>.
- [102] J. L. Leidner, “Toponym resolution in text: Annotation, evaluation and applications of spatial grounding,” *SIGIR Forum*, vol. 41, no. 2, pp. 124–126, Dec. 2007, ISSN: 0163-5840. DOI: 10.1145/1328964.1328989. [Online]. Available: <http://doi.acm.org/10.1145/1328964.1328989>.
- [103] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 9459–9474. [Online]. Available: <https://dl.acm.org/doi/pdf/10.5555/3495724.3496517>.
- [104] P. Lewis, P. Stenetorp, and S. Riedel, “Question and answer test-train overlap in open-domain question answering datasets,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 1000–1008. DOI: 10.18653/v1/2021.eacl-main.86. [Online]. Available: <https://aclanthology.org/2021.eacl-main.86>.
- [105] T. Li, A. Beirami, M. Sanjabi, and V. Smith, “Tilted empirical risk minimization,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=K5YasWXZT3O>.
- [106] Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan, “Mining evidences for named entity disambiguation,” in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’13, Chicago, Illinois, USA: ACM, 2013, pp. 1070–1078, ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2487681. [Online]. Available: <http://doi.acm.org/10.1145/2487575.2487681>.

- [107] M. D. Lieberman, H. Samet, and J. Sankaranarayanan, “Geotagging with local lexicons to build indexes for textually-specified spatial data,” in *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, Mar. 2010, pp. 201–212. DOI: 10.1109/ICDE.2010.5447903.
- [108] M. D. Lieberman and H. Samet, “Multifaceted toponym recognition for streaming news,” in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’11, Beijing, China: ACM, 2011, pp. 843–852, ISBN: 978-1-4503-0757-4. DOI: 10.1145/2009916.2010029. [Online]. Available: <http://doi.acm.org/10.1145/2009916.2010029>.
- [109] M. D. Lieberman and H. Samet, “Adaptive context features for toponym resolution in streaming news,” in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’12, Portland, Oregon, USA: ACM, 2012, pp. 731–740, ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348381. [Online]. Available: <http://doi.acm.org/10.1145/2348283.2348381>.
- [110] M. D. Lieberman, H. Samet, and J. Sankaranarayanan, “Geotagging: Using proximity, sibling, and prominence clues to understand comma groups,” in *Proceedings of the 6th Workshop on Geographic Information Retrieval*, ser. GIR ’10, Zurich, Switzerland: ACM, 2010, 6:1–6:8, ISBN: 978-1-60558-826-1. DOI: 10.1145/1722080.1722088. [Online]. Available: <http://doi.acm.org/10.1145/1722080.1722088>.
- [111] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira, “Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations,” in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2356–2362.
- [112] K. Lin, O. Tafjord, P. Clark, and M. Gardner, “Reasoning over paragraph effects in situations,” in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 58–62. DOI: 10.18653/v1/D19-5808. [Online]. Available: <https://aclanthology.org/D19-5808>.
- [113] S. Lin, J. Hilton, and O. Evans, “TruthfulQA: Measuring how models mimic human falsehoods,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229. [Online]. Available: <https://aclanthology.org/2022.acl-long.229>.
- [114] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

- [115] X. Ling, S. Singh, and D. Weld, “Design challenges for entity linking,” *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 315–328, 2015, ISSN: 2307-387X. [Online]. Available: <https://transacl.org/ojs/index.php/tacl/article/view/528>.
- [116] K. C. Litkowski, “Use of metadata for question answering and novelty tasks.,” in *TREC*, 2003, pp. 161–176.
- [117] D. Liu, Y. Gong, J. Fu, Y. Yan, J. Chen, D. Jiang, J. Lv, and N. Duan, “RikiNet: Reading Wikipedia pages for natural question answering,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 6762–6771. DOI: 10.18653/v1/2020.acl-main.604. [Online]. Available: <https://aclanthology.org/2020.acl-main.604>.
- [118] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized bert pre-training approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [119] S. Longpre, Y. Lu, Z. Tu, and C. DuBois, “An exploration of data augmentation and sampling techniques for domain-agnostic question answering,” *arXiv preprint arXiv:1912.02145*, 2019.
- [120] X. Ma, K. Sun, R. Pradeep, M. Li, and J. Lin, “Another look at DPR: Reproduction of training and replication of retrieval,” in *European Conference on Information Retrieval*, Stavanger, Norway: Springer-Verlag, 2022, pp. 613–626. DOI: 10.1007/978-3-030-99736-6_41. [Online]. Available: https://doi.org/10.1007/978-3-030-99736-6_41.
- [121] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150. [Online]. Available: <https://aclanthology.org/P11-1015>.
- [122] C. Manning, P. Raghavan, and H. Schütze, “Introduction to information retrieval,” *Natural Language Engineering*, vol. 16, no. 1, 2010.
- [123] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen, “Generation-augmented retrieval for open-domain question answering,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 4089–4100. DOI: 10.18653/v1/2021.acl-long.316. [Online]. Available: <https://aclanthology.org/2021.acl-long.316>.

- [124] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, “A SICK cure for the evaluation of compositional distributional semantic models,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 216–223. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.
- [125] K. Matsuda, A. Sasaki, N. Okazaki, and K. Inui, “Annotating geographical entities on microblog text,” in *Proceedings of The 9th Linguistic Annotation Workshop*, Denver, Colorado, USA: Association for Computational Linguistics, Jun. 2015, pp. 85–94. [Online]. Available: <http://www.aclweb.org/anthology/W15-1609>.
- [126] T. McCoy, E. Pavlick, and T. Linzen, “Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. [Online]. Available: <https://aclanthology.org/P19-1334>.
- [127] C. H. McCreery, N. Katariya, A. Kannan, M. Chablani, and X. Amatriain, “Effective transfer learning for identifying similar questions: Matching user questions to COVID-19 FAQs,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, 2020, pp. 3458–3465. DOI: 10.1145/3394486.3412861.
- [128] F. Melo and B. Martins, “Geocoding textual documents through the usage of hierarchical classifiers,” in *Proceedings of the 9th Workshop on Geographic Information Retrieval*, ser. GIR ’15, Paris, France: ACM, 2015, 7:1–7:9, ISBN: 978-1-4503-3937-7. DOI: 10.1145/2837689.2837690. [Online]. Available: <http://doi.acm.org/10.1145/2837689.2837690>.
- [129] D. Metzler and W. B. Croft, “A markov random field model for term dependencies,” in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil: Association for Computing Machinery, 2005, pp. 472–479. DOI: 10.1145/1076034.1076115.
- [130] J. Miller, K. Krauth, B. Recht, and L. Schmidt, “The effect of natural distribution shift on question answering models,” in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, PMLR, 13–18 Jul 2020, pp. 6905–6916. [Online]. Available: <https://proceedings.mlr.press/v119/miller20a.html>.
- [131] S. Min, D. Chen, H. Hajishirzi, and L. Zettlemoyer, “A discrete hard EM approach for weakly supervised question answering,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov.

- 2019, pp. 2851–2864. DOI: 10.18653/v1/D19-1284. [Online]. Available: <https://aclanthology.org/D19-1284>.
- [132] S. Min, J. Michael, H. Hajishirzi, and L. Zettlemoyer, “AmbigQA: Answering ambiguous open-domain questions,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 5783–5797. DOI: 10.18653/v1/2020.emnlp-main.466. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.466>.
- [133] R. Mirzaee, H. Rajaby Faghihi, Q. Ning, and P. Kordjamshidi, “SPARTQA: A textual question answering benchmark for spatial reasoning,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 4582–4598. DOI: 10.18653/v1/2021.naacl-main.364. [Online]. Available: <https://aclanthology.org/2021.naacl-main.364>.
- [134] L. Moncla, W. Renteria-Agualimpia, J. Nogueras-Iso, and M. Gaio, “Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus,” in *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL ’14, Dallas, Texas: ACM, 2014, pp. 183–192, ISBN: 978-1-4503-3131-9. DOI: 10.1145/2666310.2666386. [Online]. Available: <http://doi.acm.org/10.1145/2666310.2666386>.
- [135] M. Mosbach, M. Andriushchenko, and D. Klakow, “On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=nzpLWnVAyah>.
- [136] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan, “Planned protest modeling in news and social media,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI’15, Austin, Texas: AAAI Press, 2015, pp. 3920–3927, ISBN: 0-262-51129-0. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2888116.2888259>.
- [137] N. Ng, K. Cho, and M. Ghassemi, “SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1268–1283. DOI: 10.18653/v1/2020.emnlp-main.97. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.97>.
- [138] Y. Nie, S. Wang, and M. Bansal, “Revealing the importance of semantic retrieval for machine reading at scale,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong

- Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2553–2566. DOI: 10.18653/v1/D19-1258. [Online]. Available: <https://aclanthology.org/D19-1258>.
- [139] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, “Adversarial NLI: A new benchmark for natural language understanding,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4885–4901. DOI: 10.18653/v1/2020.acl-main.441. [Online]. Available: <https://aclanthology.org/2020.acl-main.441>.
- [140] T. Niven and H.-Y. Kao, “Probing neural network comprehension of natural language arguments,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4658–4664. DOI: 10.18653/v1/P19-1459. [Online]. Available: <https://aclanthology.org/P19-1459>.
- [141] R. Nogueira, W. Yang, J. Lin, and K. Cho, “Document expansion by query prediction,” *arXiv preprint arXiv:1904.08375*, 2019.
- [142] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, “Fairseq: A fast, extensible toolkit for sequence modeling,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53. DOI: 10.18653/v1/N19-4009. [Online]. Available: <https://aclanthology.org/N19-4009>.
- [143] A. Piktus, F. Petroni, V. Karpukhin, D. Okhonko, S. Broscheit, G. Izacard, P. Lewis, B. Oğuz, E. Grave, W.-t. Yih, *et al.*, “The web is your oyster—knowledge-intensive nlp against a very large web corpus,” *arXiv preprint arXiv:2112.09924*, 2021.
- [144] J. M. Ponte and W. B. Croft, “A language modeling approach to information retrieval,” in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia: Association for Computing Machinery, 1998, pp. 275–281. DOI: 10.1145/290941.291008.
- [145] Y. Qu, D. Shen, Y. Shen, S. Sajeew, W. Chen, and J. Han, “CoDA: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=Ozk9MrX1hvA>.
- [146] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang, “RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational

- Linguistics, Jun. 2021, pp. 5835–5847. DOI: 10.18653/v1/2021.naacl-main.466. [Online]. Available: <https://aclanthology.org/2021.naacl-main.466>.
- [147] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: <http://jmlr.org/papers/v21/20-074.html>.
- [148] P. Rajpurkar, R. Jia, and P. Liang, “Know what you don’t know: Unanswerable questions for SQuAD,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 784–789. DOI: 10.18653/v1/P18-2124. [Online]. Available: <https://aclanthology.org/P18-2124>.
- [149] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ questions for machine comprehension of text,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. [Online]. Available: <https://aclanthology.org/D16-1264>.
- [150] A. Rashid, V. Lioutas, and M. Rezagholizadeh, “MATE-KD: Masked adversarial TExt, a companion to knowledge distillation,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 1062–1071. DOI: 10.18653/v1/2021.acl-long.86. [Online]. Available: <https://aclanthology.org/2021.acl-long.86>.
- [151] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann, “Data augmentation can improve robustness,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: <https://openreview.net/forum?id=kgVJBBThdSZ>.
- [152] A. Ritter, Mausam, O. Etzioni, and S. Clark, “Open domain event extraction from twitter,” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’12, Beijing, China: ACM, 2012, pp. 1104–1112, ISBN: 978-1-4503-1462-6. DOI: 10.1145/2339530.2339704. [Online]. Available: <http://doi.acm.org/10.1145/2339530.2339704>.
- [153] A. Roberts, C. Raffel, and N. Shazeer, “How much knowledge can you pack into the parameters of a language model?” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 5418–5426. DOI: 10.18653/v1/2020.emnlp-main.437. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.437>.

- [154] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, *et al.*, “Okapi at trec-3,” *Nist Special Publication Sp*, vol. 109, 1995.
- [155] A. Rogers, “Changing the world by changing the data,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 2182–2194. DOI: 10.18653/v1/2021.acl-long.170. [Online]. Available: <https://aclanthology.org/2021.acl-long.170>.
- [156] A. Rogers, M. Gardner, and I. Augenstein, “QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension,” *arXiv preprint arXiv:2107.12708*, 2021.
- [157] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, “Everyone wants to do the model work, not the data work: Data cascades in high-stakes ai,” in *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan: Association for Computing Machinery, 2021, pp. 1–15. DOI: 10.1145/3411764.3445518.
- [158] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,” vol. arXiv:1910.01108, 2019. [Online]. Available: <http://arxiv.org/abs/1910.01108>.
- [159] J. Santos, I. Anastácio, and B. Martins, “Using machine learning methods for disambiguating place references in textual documents,” *GeoJournal*, vol. 80, no. 3, pp. 375–392, Jun. 2015, ISSN: 1572-9893. DOI: 10.1007/s10708-014-9553-y. [Online]. Available: <https://doi.org/10.1007/s10708-014-9553-y>.
- [160] R. M. Schwarcz, J. F. Burger, and R. F. Simmons, “A deductive question-answerer for natural language inference,” *Communications of the ACM*, vol. 13, no. 3, pp. 167–183, Mar. 1970. DOI: 10.1145/362052.362058.
- [161] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96. DOI: 10.18653/v1/P16-1009. [Online]. Available: <https://aclanthology.org/P16-1009>.
- [162] M. Seo, J. Lee, T. Kwiatkowski, A. Parikh, A. Farhadi, and H. Hajishirzi, “Real-time open-domain question answering with dense-sparse phrase index,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4430–4441. DOI: 10.18653/v1/P19-1436. [Online]. Available: <https://aclanthology.org/P19-1436>.

- [163] S. Shakeri, C. Nogueira dos Santos, H. Zhu, P. Ng, F. Nan, Z. Wang, R. Nallapati, and B. Xiang, “End-to-end synthetic data generation for domain adaptation of question answering systems,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 5445–5460. DOI: 10.18653/v1/2020.emnlp-main.439. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.439>.
- [164] D. Shen, M. Zheng, Y. Shen, Y. Qu, and W. Chen, “A simple but tough-to-beat data augmentation approach for natural language understanding and generation,” *arXiv preprint arXiv:2009.13818*, 2020.
- [165] W. Shen, J. Wang, and J. Han, “Entity linking with a knowledge base: Issues, techniques, and solutions,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 443–460, 2015, ISSN: 1041-4347. DOI: 10.1109/TKDE.2014.2327028.
- [166] L. Shi and J.-Y. Nie, “Using various term dependencies according to their utilities,” in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, 2010, pp. 1493–1496. DOI: 10.1145/1871437.1871655.
- [167] M. Shirakawa, T. Hara, and S. Nishio, “N-gram IDF: A global term weighting scheme based on information distance,” in *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy: International World Wide Web Conferences Steering Committee, 2015, pp. 960–970. DOI: 10.1145/2736277.2741628.
- [168] R. F. Simmons, “Natural language question-answering systems: 1969,” *Commun. ACM*, vol. 13, no. 1, pp. 15–30, Jan. 1970. DOI: 10.1145/361953.361963.
- [169] R. F. Simmons, S. Klein, and K. McConlogue, “Indexing and dependency logic for answering english questions,” *American Documentation*, vol. 15, no. 3, pp. 196–204, 1964. DOI: 10.1002/asi.5090150306.
- [170] D. Singh, S. Reddy, W. Hamilton, C. Dyer, and D. Yogatama, “End-to-end training of multi-document reader and retriever for open-domain question answering,” in *Advances in Neural Information Processing Systems*, vol. 34, Curran Associates, Inc., 2021, pp. 25 968–25 981. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/file/da3fde159d754a2555eaa198d2d105b2-Paper.pdf>.
- [171] S. K. Singh and D. Rafiei, “Strategies for geographical scoping and improving a gazetteer,” in *Proceedings of the 2018 World Wide Web Conference*, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 1663–1672. DOI: 10.1145/3178876.3186078.

- [172] P. Slavík, “A tight analysis of the greedy algorithm for set cover,” in *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, ser. STOC '96, Philadelphia, Pennsylvania, USA: ACM, 1996, pp. 435–441, ISBN: 0-89791-785-5. DOI: 10.1145/237814.237991. [Online]. Available: <http://doi.acm.org/10.1145/237814.237991>.
- [173] M. Speriosu and J. Baldrige, “Text-driven toponym resolution using indirect supervision,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 1466–1476. [Online]. Available: <http://www.aclweb.org/anthology/P13-1144>.
- [174] A. Spitz, J. Geiß, and M. Gertz, “So far away and yet so close: Augmenting toponym disambiguation and similarity with text-based networks,” in *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data*, ser. GeoRich '16, San Francisco, California: ACM, 2016, 2:1–2:6, ISBN: 978-1-4503-4309-1. DOI: 10.1145/2948649.2948651. [Online]. Available: <http://doi.acm.org/10.1145/2948649.2948651>.
- [175] M. Srikanth and R. Srihari, “Incorporating query term dependencies in language models for document retrieval,” in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada: Association for Computing Machinery, 2003, pp. 405–406. DOI: 10.1145/860435.860523.
- [176] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang, “Mitigating gender bias in natural language processing: Literature review,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1630–1640. DOI: 10.18653/v1/P19-1159. [Online]. Available: <https://aclanthology.org/P19-1159>.
- [177] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://openreview.net/forum?id=wCu6T5xFjeJ>.
- [178] A. Trotman, A. Puurula, and B. Burgess, “Improvements to BM25 and language models examined,” in *Proceedings of the 2014 Australasian Document Computing Symposium*, 2014, pp. 58–65.
- [179] R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese, “Generalizing to unseen domains via adversarial data augmentation,” *Advances in neural information processing systems*, vol. 31, 2018. [Online]. Available: <https://dl.acm.org/doi/pdf/10.5555/3327345.3327439>.
- [180] E. M. Voorhees, “The TREC-8 question answering track report.,” in *TREC*, vol. 99, 1999, pp. 77–82.

- [181] H. Wan, S. Feng, C. Gunasekara, S. S. Patel, S. Joshi, and L. Lastras, “Does structure matter? encoding documents for machine reading comprehension,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 4626–4634. DOI: 10.18653/v1/2021.naacl-main.367. [Online]. Available: <https://aclanthology.org/2021.naacl-main.367>.
- [182] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “GLUE: A multi-task benchmark and analysis platform for natural language understanding,” in *International Conference on Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=rJ4km2R5t7>.
- [183] D. Wang, Y. Li, L. Wang, and B. Gong, “Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1498–1507.
- [184] S. Wang, M. Yu, X. Guo, Z. Wang, T. Klinger, W. Zhang, S. Chang, G. Tesauro, B. Zhou, and J. Jiang, “R³: Reinforced ranker-reader for open-domain question answering,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, Apr. 2018. DOI: 10.1609/aaai.v32i1.12053.
- [185] Z. Wang, P. Ng, X. Ma, R. Nallapati, and B. Xiang, “Multi-passage BERT: A globally normalized BERT model for open-domain question answering,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5878–5882. DOI: 10.18653/v1/D19-1599. [Online]. Available: <https://aclanthology.org/D19-1599>.
- [186] J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6382–6388. DOI: 10.18653/v1/D19-1670. [Online]. Available: <https://aclanthology.org/D19-1670>.
- [187] A. Williams, N. Nangia, and S. Bowman, “A broad-coverage challenge corpus for sentence understanding through inference,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. [Online]. Available: <https://aclanthology.org/N18-1101>.

- [188] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. [Online]. Available: <https://aclanthology.org/2020.emnlp-demos.6>.
- [189] W. A. Woods, “Semantics and quantification in natural language question answering,” in, ser. *Advances in Computers*, M. C. Yovits, Ed., vol. 17, Elsevier, 1978, pp. 1–87. DOI: 10.1016/S0065-2458(08)60390-3.
- [190] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, “Conditional BERT contextual augmentation,” in *International Conference on Computational Science*, Springer International Publishing, 2019, pp. 84–95.
- [191] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, “Unsupervised data augmentation for consistency training,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6256–6268. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf>.
- [192] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. N. Bennett, J. Ahmed, and A. Overwijk, “Approximate nearest neighbor negative contrastive learning for dense text retrieval,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=zeFrfgYzln>.
- [193] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky, “Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy,” in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, Barcelona, Spain: Association for Computing Machinery, 2020, pp. 547–558. DOI: 10.1145/3351095.3375709.
- [194] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin, “End-to-end open-domain question answering with BERTserini,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 72–77. DOI: 10.18653/v1/N19-4013. [Online]. Available: <https://aclanthology.org/N19-4013>.
- [195] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, “HotpotQA: A dataset for diverse, explainable multi-hop question answering,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2369–2380. DOI: 10.18653/v1/D18-1259. [Online]. Available: <https://aclanthology.org/D18-1259>.

- [196] M. Yi, L. Hou, L. Shang, X. Jiang, Q. Liu, and Z.-M. Ma, “Reweighting augmented samples by minimizing the maximal expected loss,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=9G5MIc-goqB>.
- [197] J. Yu and D. Rafiei, “Geotagging named entities in news and online documents,” in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ser. CIKM ’16, Indianapolis, Indiana, USA: ACM, 2016, pp. 1321–1330, ISBN: 978-1-4503-4073-1. DOI: 10.1145/2983323.2983795. [Online]. Available: <http://doi.acm.org/10.1145/2983323.2983795>.
- [198] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, “HellaSwag: Can a machine really finish your sentence?” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4791–4800. DOI: 10.18653/v1/P19-1472. [Online]. Available: <https://aclanthology.org/P19-1472>.
- [199] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=r1Ddp1-Rb>.
- [200] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in *Advances in Neural Information Processing Systems*, vol. 28, Curran Associates, Inc., 2015, pp. 649–657. [Online]. Available: <https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf>.
- [201] Y. Zhang, J. Baldridge, and L. He, “PAWS: Paraphrase adversaries from word scrambling,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1298–1308. DOI: 10.18653/v1/N19-1131. [Online]. Available: <https://aclanthology.org/N19-1131>.
- [202] B. Zheng, H. Wen, Y. Liang, N. Duan, W. Che, D. Jiang, M. Zhou, and T. Liu, “Document modeling with graph attention networks for multi-grained machine reading comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 6708–6718. DOI: 10.18653/v1/2020.acl-main.599. [Online]. Available: <https://aclanthology.org/2020.acl-main.599>.
- [203] G. Zhou, T. He, J. Zhao, and P. Hu, “Learning continuous word embedding with metadata for question retrieval in community question answering,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Com-

putational Linguistics, Jul. 2015, pp. 250–259. DOI: 10.3115/v1/P15-1025.
[Online]. Available: <https://aclanthology.org/P15-1025>.